PROBABILIDAD Y ESTADÍSTICA

PARA INGENIERÍA Y CIENCIAS

Israel Mallma Perez
Lita Salazar Vázquez
Erik Mallma Perez

Huancayo, Junín, Perú

PROBABILIDAD Y ESTADÍSTICA

PARA INGENIERÍA Y CIENCIAS

Autores

Israel Mallma Perez
Lita Salazar Vázquez
Erik Mallma Perez

Editor Israel Mallma

Huancayo, Junín, Perú

PROBABILIDAD Y ESTADÍSTICA PARA INGENIERÍA Y CIENCIAS

Autores:

© Israel Jimmy Mallma Perez

© Lita Pierina Salazar Vázquez

© Erik Edgar Mallma Perez

Editado por:

© Israel Jimmy Mallma Perez

Jirón La Mar 264 Huancayo

Cel. 964349461

mpisrael200@gmail.com Huancayo – Perú

Primera edición, junio 2019

Tiraje: Ninguna (edición electrónica)

ISBN: 978-612-00-4450-6

ISBN: 978-612-00-4450-6

9 786120 044506

Libro electrónico disponible en:

https://sites.google.com/view/israelmallma/p%C3%A1gina-principal

PROLOGO

El objetivo principal perseguido con este libro es ofrecer a los estudiantes un manual estadístico básico que, sin descuidar el rigor conceptual, proporciona una visión práctica e intuitiva de la estadística descriptiva y el cálculo de probabilidades, campos básico y fundamentos de la ciencia estadística. El contenido del manual estadística descriptiva y probabilidad se organizó en dos partes, en la primera descriptiva estadística se estudia, siendo la segunda dedicada al cálculo de las probabilidades. En cualquier caso, se debe notar que, en un análisis inferencial, también debemos hacer un estudio descriptivo de la muestra. La importancia del estudio multidimensional se debe al hecho de considerar la existencia de posibles interacciones entre las diferentes variables en estudio.

Cada uno de los capítulos del libro comienza con una presentación del tipo de problema a ser resuelto, la exposición continua de las ilustradas con varios ejemplos contenidos y, a veces acompañada de un ejercicio que busca profundizar un tema de interés, terminar con ejercicios resueltos, que intenta globalizar los aspectos más relevantes del capítulo. El manual, como la actual, es salpicado con ejemplos y ejercicios hecho y, sin la pretensión de ser un libro teórico, el lector que así lo desee puede encontrar respuestas fundamentadas cuestiones conceptuales que están surgiendo.

ÍNDICE

PROLOGO 4
HISTORIA DE LA ESTADÍSTICA 12
FUNDAMENTOS DE LA ESTADÍSTICA 14
CONCEPTOS BÁSICOS DE LA ESTADÍSTICA 14
DIVISIÓN DE LA ESTADÍSTICA 15
1. INTRODUCCIÓN AL ANÁLISIS DE DATOS 16
1.1. La probabilidad en la actualidad 16
1.2. Análisis de tendencias estadísticas 16
1.2.1. Medidas de tendencia central 17
1.2.2. Desviaciones 17
1.2.3. Medidas de dispersión 18
1.2.4. Medidas de tendencia central y de dispersión en datos agrupados 19
1.2.5. Distribución muestral de medias 23
1.3. Datos discretos y continuos 25
1.3.1. Los dos principales tipos de datos: cualitativos y cuantitativos 25
1.3.2. Tipos cuantitativos: datos continuos y datos discretos 26
1.3.3. Tipos Cualitativos: Datos Binomiales, Datos Nominales y Datos Ordinales. 26
1.3.4. Modelado estadístico, inspección científica y representación grafica de datos 27
2. PROBABILIDAD 37
2.1. Espacio muestral 37
2.1.1. Tipos de espacio muestral 38
2.2. Eventos 39
2.2.1. Tipos de eventos 40
2.3. Conteo de puntos muestrales 40
2.3.1. Regla MNT 41
2.3.2. Combinaciones 41
2.3.3. Permutaciones 45
2.4. Probabilidad de un evento 47
2.4.1. Cálculo o medición de la Probabilidad 47
2.4.2. Condiciones importantes 49
2.5. Reglas aditivas 49
2.5.1. Eventos no Mutuamente Excluyentes 49
2.5.2. Eventos Mutuamente Excluyentes 50

2.5.3. Eventos Complementarios 50
2.6. Probabilidad condicional 51
2.7. Reglas multiplicativas 52
2.7.1. Eventos independientes 52
2.8. Regla de Bayes 52
2.9. Ejercidos de Repaso 54
3. VARIABLES ALEATORIAS Y DISTRIBUCIONES DE PROBABILIDAD 56
3.1. Concepto de variable aleatoria 56
3.2. Distribuciones discretas de probabilidad 58
3.2.1. La media y la varianza de las distribuciones discretas 59
3.3. Distribuciones continuas de probabilidad 60
3.3.1. Características: 60
3.3.2. Calculo de media y desviación estándar para una distribución continua 60
3.4. Distribuciones empíricas 63
3.5. Distribuciones de probabilidad conjunta 65
4. ESPERANZA MATEMÁTICA 65
4.1. Media de una variable aleatoria 65
4.1.1. Variables aleatorias discretas 65
4.1.2. Parámetros de una variable aleatoria 66
4.2. Varianza y covarianza 66
3.5.1. La Varianza 66
3.5.2. Covarianza 69
4.3. Medias y varianzas de combinaciones lineales de variables aleatorias 71
4.4. Teorema de Chebyshev 72
4.4.1. Desigualdad de Chebyshev 72
5. DISTRIBUCIÓN DE PROBABILIDAD DISCRETA 73
5.1. conceptos Generales 73
5.2. distribución uniforme discreta 75
5.2.1. Distribución Uniforme discreta (a,b) 75
5.2.2. Distribución Binomial (n,p) 76
5.2.3. Distribución Hipergeométrica (N,R,n) 77
5.2.4. Distribución Geométrica (p) 78
5.2.5. Distribución Binomial negativa (r,p) 79
5.2.6. Distribución Poisson (lambda) 80
5.2.7. Resumen 83

6. ALGUNAS DISTRIBUCIONES CONTINUAS DE PROBABILIDAD 85
6.1. Distribución Uniforme (a,b) 85
6.2. Distribución Normal (Mu, Sigma) 86
6.3. Distribución Lognormal (Mu, Sigma) 88
6.4. Distribución Logística (a, b) 89
6.5. Distribución Beta (p,q) 90
6.6. Distribución Gamma (a,p) 91
6.7. Distribución Exponencial (lambda) 93
6.8. Distribución Ji-cuadrado (n) 94
6.9. Distribución t de Student (n) 95
6.10. Distribución F de Snedecor (n,m) 96
7. FUNCIONES DE VARIABLES ALEATORIAS 98
7.1. Funciones de variables aleatorias cambio de variable 98
7.1.1. Teorema 98
7.1.2. Teorema general de Cambio de Variable 98
7.2. Cambio de variable discreto 99
7.2.1. Teorema 99
7.3. Cambio De Variable Continuo 99
7.3.1. Teorema de cambio de variable de continuo a discreto 99
7.3.2. Teorema de cambio de variable continuo a continuo 100
8. DISTRIBUCIONES FUNDAMENTALES DE MUESTREO Y DESCRIPCIONES DE DATOS 101
8.1. Muestreo aleatorio 101
8.1.1. Ventajas del muestreo aleatorio simple 101
8.1.2. Desventajas del muestreo aleatorio simple 101
8.2. Inferencia de la media poblacional a partir de la media muestral 102
8.2.1. Inferencia Paramétrica 102
8.2.2. Inferencia No Parámetrica 102
8.2.3. Estimación 103
8.2.4. Teorema del limite central 103
8.2.5. Muestreo 104
8.1. distribuciones muestrales 106
8.2. distribuciones muestrales de medias 108
8.3. Distribución muestral de s al cuadrado 109
8.2.6. Definiciones: 109

8.4. distribución T 110
8.5. distribución F..... 112
9. PROBLEMAS DE ESTIMACIÓN DE UNA Y DOS MUESTRAS..... 113
9.1. introducción..... 113
9.2. Inferencia estadística 113
9.3. Métodos clásicos de estimación 114
9.3.1. Estimación puntual 114
9.3.2. Estimación por intervalos 116
9.4. Una sola muestra estimación de la media 119
9.5. Error estándar de una estimación puntual 123
9.6. Intervalos de tolerancia 124
9.7. Límites de tolerancia 125
9.8. Dos muestras estimación de la diferencia entre 2 medias 126
9.9. Observaciones pareadas 127
9.10. Una sola muestra estimación de una proporción..... 129
9.11. Dos muestras estimación de la diferencia entre dos proporciones..... 130
9.12. una sola muestra estimación de la varianza 131
9.13. dos muestras estimación de la proporcion de dos varianzas 132
9.14. Estimación de la máxima verosimilitud (opcional)..... 134
9.15. Posibles riesgos y errores conceptuales: relación con el material de otros capítulos 136
10. PRUEBAS DE HIPÓ TESIS DE UNA Y DOS MUESTRAS..... 137
10.1. Hipó tesis estadística 137
10.2. Prueba de una hipó tesis estadística 138
10.3. Pruebas de una cola y dos colas 147
10.4. El uso de los valores P en la toma de decisiones 149
10.5. Pruebas relacionadas con una sola media (variancia conocida)..... 152
10.6. Relación con la estimación del intervalo de confianza 154
10.7. Pruebas relacionadas con una sola media (variancia desconocida) 155
10.8. Pruebas sobre dos medias (Dos muestras) 156
10.9. Selección del tamaño de la muestra para la prueba de medias..... 158
10.10. Pruebas relacionadas con proporciones 162
10.11. Pruebas de una y dos muestras referentes a varianzas..... 164
10.12. Prueba de la bondad de ajuste..... 166
10.13. Prueba de la independencia (datos categóricos) 169

10.14. Prueba de Homogeneidad 171
11. REGRESIÓN LINEAL SIMPLE Y CORRELACIÓN 172
11.1. Introducción a la regresión lineal 172
11.2. regresión lineal simple 173
11.3. Propiedades de los estimadores de mínimos cuadrados 177
11.4. inferencias acerca de los coeficientes de regresión 181
11.5. Predicción 183
11.6. elección de un modelo de regresión 185
11.7. procedimiento del análisis de varianza 186
11.8. Prueba de la linealidad de la regresión datos con observaciones repetidas 187
11.9. Gráficas y transformaciones de datos 193
11.10. estudios de regresión lineal simple 194
11.11. Correlación 196
12. REGRESIÓN LINEAL MÚLTIPLE 199
12.1. introducción 199
12.2. estimación de coeficientes 200
12.3. modelo de regresión lineal con el uso de matrices 201
12.4. propiedades de los estimadores de mínimos cuadrados 203
12.5. inferencias en la regresión lineal múltiple 205
12.6. elección de un modelo de ajuste la prueba de hipótesis 207
12.7. Caso especial de ortogonalidad 209
12.8. métodos secuenciales para la Selección del modelo 211
12.9. estudio de residuos y violación de suposiciones 212
12.10. validación Cruzada C y otros criterios para la Selección del modelo 216
13. EXPERIMENTOS DE UN FACTOR GENERAL 221
13.1. técnica de análisis de varianza 221
13.2. la estrategia del diseño experimental 223
13.3. análisis de varianza unilateral diseño completamente aleatorio 223
13.4. pruebas de la igualdad de varias varianzas 226
13.5. Comparaciones de un solo grado de libertad 227
13.6. comparaciones múltiples 228
13.7. comparación de tratamientos con un control 228
13.8. Comparación de un conjunto de tratamientos en bloques 232
13.9. Diseño de bloques completos aleatorios 233
13.10. Métodos gráficos y diagnósticos posteriores 235

13.11. Transformaciones de datos en el análisis de varianza 236
13.12. modelos de efectos aleatorios 236
13.13. Estudio de caso 239
14. EXPERIMENTOS FACTORIALES 241
14.1. introducción 241
14.2. interacción y el experimento de dos factores 241
14.3. análisis de varianza de dos factores 244
14.4. Análisis gráfico en el problema de dos factores 248
14.5. Experimentos de tres factores 250
14.6. Experimentos factoriales II y III 254
14.7. Posibles riesgos y errores conceptuales; relación con el material de otros capítulos 255
15. EXPERIMENTOS 2 FACTORIALES Y FRACCIONES 256
15.1. Introducción 256
15.2. Análisis de varianza 256
15.3. Experimento 2 K factorial sin replicas 259
15.4. Experimentos factoriales de bloque sincompletos 264
15.5. Diseño ortogonal 266
15.6. Experimentos factoriales fraccionados 268
15.7. Análisis de Experimentos factoriales fraccionados 272
15.8. Diseños de fracciones superiores y depurados 272
15.9. Construcción de diseños de resolución III y IV con 8 16 y 32 puntos de diseño 273
15.10. Otros diseños de resolución III de dos niveles diseños de Plackett-Burman 275
15.11. Diseño robusto de parámetros de Taguchi 276
16. ESTADÍSTICA NO PARAMÉTRICA 277
16.1. Pruebas no paramétricas 277
16.2. Prueba de signo 278
16.3. Prueba de rango con signo 280
16.4. Prueba de la suma de rangos 282
16.5. Prueba de Kruskal Wallis 283
16.6. Pruebas de rachas 285
16.7. Límites de tolerancia 286
16.8. Coeficiente de correlación de rango 287
17. CONTROL ESTADÍSTICO DE CALIDAD 288
17.2. Introducción 288

17.3. Naturaleza de los límites de control 289
17.4. Propósitos de la grafica de control 289
17.5. Graficas de control para las variables 290
17.6. Graficas de control para artículos defectuosos (uso del modelo de Poisson) .. 292
17.7. Graficas de control de cusum 295
BIBLIOGRAFÍA 297
APÉNDICE A TABLAS Y DEMOSTRACIONES ESTADÍSTICAS..... 301

HISTORIA DE LA ESTADÍSTICA

Hablar de la historia de la estadística, implica hablar del inicio de la civilización. Existe una gran diversidad sobre los inicios de la estadística. En este texto, haremos un recorrido breve sobre el desarrollo de la Estadística.

El término Estadística fue empleado por primera vez por Gottfried Achenwall (1719-1772), el cual proviene de la palabra de origen italiana "Statistik".

La estadística fue utilizada por los egipcios para analizar los datos de la población, la renta (Aproximadamente siglo XXXII A. de C). En el antiguo testamento, primera parte de la Biblia, se encuentran los libros de "Números" y "Crónicas", donde se pueden observar algunos registros estadísticos.

El libro de Números contiene dos censos de la población de Israel y el de Crónicas describe la prosperidad económica de las tribus judías.

En sentido general, los grandes registros estadísticos de la antigüedad descansan en censos y cobro de impuestos.

Se registros de que el Imperio romano fue el primero en recopilar gran cantidad de datos sobre la población, superficie y la renta de todas las comarcas bajo su dominio.

En 1662 se conoce como el primer estudio estadístico de mayor trascendencia sobre población, el trabajo titulado "Observations on the London Bills of Mortality" (Londres).

Para hablar sobre el desarrollo de la estadística, hay que destacar los aportes de grandes personalidades de la ciencia. A continuación, se presentan algunas de estas grandes personalidades:

John Graunt (1620-1674): Es el primer estadístico que se va a considerar como el padre de la Estadística. En 1662 publicó "Observaciones naturales y políticas", fundamentado en los boletines de mortalidad, considerándose este documento como el punto de partida de la estadística.

Thomas Bayes (1702-1761): Sus actividades científicas, fueron las matemáticas y la inferencia estadística. Después de la muerte de Bayes, Price, recibe de parte de los familiares de Bayes, escritos sobre matemáticas que Bayes había dejado, para que Price los estudiara. El trabajo de Bayes demuestra la divergencia de la serie log(n!), corrigiendo a la de Moivre. El ensayo Bayes determina un intervalo de confianza bayesiano para el parámetro &μ de una distribución de Bernoulli a partir de n repeticiones del experimento de Bernoulli.

Laplace (1749-1827): Pierre Simon de Laplace nace en Beaumont-en-Auge. En 1774 publica Memoria sobre la probabilidad de las causas por los sucesos.

Karl Pearson (1857-1936): En 1905 publicó un artículo titulado Sobre la teoría general de la asimetría, la correlación y la regresión no lineal. Algunas contribuciones de K. Pearson a la Teoría de la Probabilidad y a la Inferencia Estadística son:

Introduce su familia de curvas y ajusta sus parámetros, introduciendo el método de los momentos. Esta familia de curvas son las soluciones de una ecuación diferencial e incluye a las distribuciones, Beta asimétrica, la Beta simétrica, la Gamma y la Normal, entre otras.

Define el coeficiente de correlación lineal precisando las ideas introducidas por Galton.

William Sealy Gosset (STUDENT), (1876-1937): Publica trabajos sobre control de calidad, la ley de errores, la distribución de los errores.

Algunas contribuciones de (Student) William Sealy Gosset son:

La demostración de la convergencia de la distribución binomial a la distribución de Poisson.

Estableciendo su distribución, donde X es la media muestral y S2, la cuasi varianza muestral.

Introdujo la función de potencia de un contraste dado por la región crítica.

Ronald Aylmer Fisher (1890-1962): Escribió "La Fundamentación Matemática de la Estadística Teórica, introduciendo la noción de modelo estadístico y los conceptos de consistencia, eficiencia, precisión, validación e información. Desarrolló el análisis de la varianza y los principios del diseño de experimentos. En 1925 publica su primer libro titulado "Métodos Estadísticos". En 1930 formula La Teoría Genética de la Selección Natural. En 1956 publica Métodos Estadísticos e Inferencia Científica.

FUNDAMENTOS DE LA ESTADÍSTICA

La estadística recoge, ordena y analiza datos para estudiar las características o el comportamiento de un colectivo.

Es el conjunto de métodos científicos ligados a la toma, organización, recopilación, presentación y análisis de datos, tanto para la deducción de conclusiones como para tomar decisiones razonables de acuerdo con el análisis.

Sobre estadística se presentan diversos conceptos; por tal razón les presento los siguientes conceptos sobre estadística:

- La estadística es una rama de las matemáticas que se ocupa de reunir, organizar y analizar datos numéricos y que ayuda a resolver problemas como el diseño de experimentos y la toma de decisiones.

- La Estadística es la parte de las Matemáticas que se encarga de recoger datos, organizarlos, tabularlos, representarlos gráficamente, interpretarlos y analizarlos, para sacar conclusiones, tomar decisiones o predecir, a partir del estudio o investigación que se esté realizando sobre alguna característica, propiedad o situación en una población determinada.

CONCEPTOS BÁSICOS DE LA ESTADÍSTICA

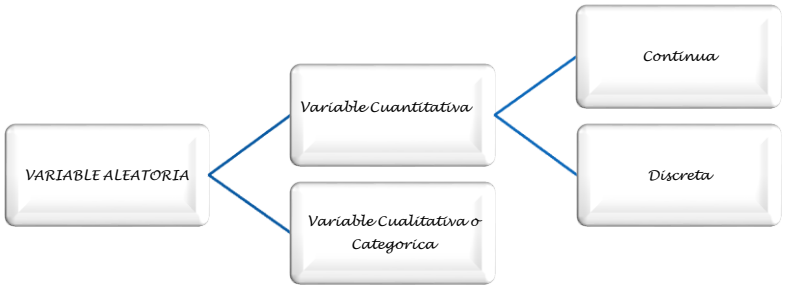
Daremos a conocer solamente el vocabulario estadístico básico, los demás términos se definirán ha mediado que se vaya presentando la necesidad.

VARIABLE

Es el conjunto de las características de las entidades que interesan en una investigación estadística.

Los valores que pueden tomar una variable en estadística no se pueden medir antes de realizar el experimento estadístico; por lo tanto el valor es aleatorio(al azar).

Las características que se pueden medir en una “entidad” son de tipo cuantitativo (categórico); como cada característica es una variable aleatoria, estas tienen las siguientes subdivisiones:



VARIABLE CUANTITATIVA

Se dice que una variable es cuantitativa siempre que los valores que son el resultado de medidas numéricas. Ejemplo de variable cuantitativa son la densidad, el peso, la viscosidad, la temperatura, etc.

Variable Cuantitativa Continua

Una variable continua es aquella que teóricamente pueden tomar cualquier valor dentro de un intervalo de valores. Es decir, una variable continua se mede uniformemente. Por ejemplo, si la variable en gramos la definimos como X, esta puede estar en el intervalo $50g \leq X \leq 70g$.

Cuantitativa Discreta Variable

Cuando los valores numéricos que pueden tomar una variable son números enteros, la variable se denomina discreta. Por ejemplo, si la variable es la edad cumplida en años de una persona, esta solo puede tomar los valores de 1,2, 3,…, 120 (si la expectativa de vida es 120 años)

VARIABLE CUALITATIVA

Hay muchas cosas en que no es posible hacer medidas numéricas. Por ejemplo, la variable "color" puede recibir los valores cualitativos de rojo, verde, amarillo, etc. Una variable cuyos valores consiste en categorías de clasificación se denomina variable cualitativa.

DIVISIÓN DE LA ESTADÍSTICA

La estadística se divide en dos grandes partes:

Estadística descriptiva: Se encarga de la recolección, descripción, visualización y resumen de los datos, y que se pueden presentar de forma numérica o de forma gráfica.

La Estadística descriptiva, se encarga del "estudio sobre la población completa, observando una característica de la misma y calculando unos parámetros que den información global de toda la población".

- ✓ Para variables cualitativas se calcula los porcentajes y las tasas
- ✓ Para variables cuantitativas se tiene las medidas de tendencia central (promedios, moda mediana) y las medidas de dispersión (dispersión varianza rango entre otros)

Estadística inferencial: Se encarga de las predicciones relacionadas a los fenómenos estudiados, considerando la selección de tipo aleatoria y la incertidumbre en las observaciones.

La Estadística inferencial, "realiza el estudio descriptivo sobre un subconjunto de la población llamado muestra y, posteriormente, extiende los resultados obtenidos a toda la población".

Para variables cualitativas: el chi cuadrado

Para variables cualitativas: la paramétrica (T de Student, ANOVA) para las no paramétricas (U de Mann-Whitney, Prueba de Kruskal-Wallis) y para Asociaciones de variables la regresión y correlación.

1. INTRODUCCIÓN AL ANÁLISIS DE DATOS

1.1. La probabilidad en la actualidad

. La importancia de la probabilidad reside en el hecho de que, por medio de ese recurso matemático, es posible ajustar con la mayor precisión posible los imponderables debidos al azar en los más variados campos de la ciencia y de la vida cotidiana. En efecto, la probabilidad es una estrategia por la cual tratamos de estimar la frecuencia con que un determinado resultado se obtiene en el contexto de una experiencia en la que se conocen todos los resultados posibles. Así, el ejemplo más tradicional es definir cuál es la prevalencia de obtener un número jugando un dado. En seis resultados posibles, sólo es posible obtener un número cada vez que se contabiliza el dato.

En este caso, la probabilidad se puede expresar como uno en seis, un sexto, la sexta parte o, en términos matemáticos precisos, 0,16 o 16%. La importancia esencial de la aplicación de métodos de cálculo de probabilidad reside en su capacidad para estimar o prever eventos. Cuanto mayor sea la cantidad de datos disponibles para calcular la probabilidad de un evento, más preciso será el resultado calculado. Dada la complejidad de los sistemas en que la teoría de la probabilidad es usualmente aplicada, son necesarios modelos computacionales y estadísticos de gran elaboración, lo que sería imposible no contar con los recursos tecnológicos modernos relacionados a la computación.

Un buen ejemplo de su aplicabilidad diaria es el análisis del comercio de commodities en las actuales relaciones internacionales. Como la mayoría de los factores involucrados en la estimación de la producción es aleatoria, la teoría de la probabilidad es de gran importancia, ya que intenta ajustar en conceptos matemáticos cuál será la evolución de los eventos para calcular, por ejemplo, la producción final. de cereales, combustibles fósiles y otros recursos de una zona geográfica. Por lo tanto, la probabilidad es una herramienta fundamental en la planificación estratégica de los movimientos sociales, económicos y laborales de toda la comunidad. El análisis permite describir los comportamientos de la información, obtener conclusiones y dar recomendaciones para la toma de decisiones.

Algunos fenómenos poseen características aleatorias, por lo tanto, es conveniente estudiar la posibilidad de ocurrencia de eventos relacionados a este fenómeno, esa posibilidad será denominada probabilidad. Con su estudio, los resultados de los fenómenos o experimentos aleatorios son analizados y con ellos es posible tomar decisiones en la presencia de incertidumbre.

1.2. Análisis de tendencias estadísticas

Las medidas de tendencia central son medidas estadísticas que pretenden resumir en un solo valor a un conjunto de valores. Representan un centro en torno al cual se encuentra ubicado el conjunto de los datos. Dicho en otros términos las medidas de dispersión pretenden evaluar en qué medida los datos difieren entre sí. De esta forma, ambos tipos de medidas usadas en conjunto permiten describir un conjunto de datos entregando información acerca de su posición y su dispersión.

Los procedimientos para obtener las medidas estadísticas difieren levemente dependiendo de la forma en que se encuentren los datos.

1.2.1. Medidas de tendencia central

A. Promedio o media

La medida de tendencia central más conocida y utilizada es la media aritmética o promedio aritmético. Se representa por la letra griega μ cuando se trata del promedio del universo o población y por Ȳ (léase Y barra) cuando se trata del promedio de la muestra.

Si una muestra tiene cuatro observaciones: 6, 5, 2 y 2, por definición el estadígrafo será:

$$\bar{y} = \frac{6+5+2+2}{4} = 3$$

En este caso el promedio está dado por

$$\bar{y} = \frac{y_1 + y_2 + y_3 + y_4 \dots\dots\dots . y_n}{n}$$

De aquí se desprende la fórmula definitiva del promedio:

$$\bar{Y} = \frac{\sum_1^n Y_1}{n}$$

B. Mediana

Otra medida de tendencia central es la mediana. La mediana es el valor de la variable que ocupa la posición central, cuando los datos se disponen en orden de magnitud. Es decir, el 50% de las observaciones tiene valores iguales o inferiores a la mediana y el otro 50% tiene valores iguales o superiores a la mediana.

Si el número de observaciones es par, la mediana corresponde al promedio de los dos valores centrales. Por ejemplo, en la muestra 3, 9, 11, 15, la mediana es (9+11)/2=10.

C. Moda

La moda de una distribución se define como el valor de la variable que más se repite. En un polígono de frecuencia la moda corresponde al valor de la variable que está bajo el punto más alto del gráfico. Una muestra puede tener más de una moda.

1.2.2. Desviaciones

Se define como la desviación de un dato a la diferencia entre el valor del dato y la media:

$$\text{Desviación} = Y_i - \bar{Y}$$

Ejemplo de desviaciones:

Y_i	$\bar{Y}$	$Y_i - \bar{Y}$
3	3	0
5	3	+2
2	3	-1
2	3	-1
Suma		0

Una propiedad interesante de la media aritmética es que la suma de las desviaciones es cero.

1.2.3. Medidas de dispersión

Las medidas de dispersión entregan información sobre la variación de la variable. Pretenden resumir en un solo valor la dispersión que tiene un conjunto de datos. Las medidas de dispersión más utilizadas son: Rango de variación, Varianza, Desviación estándar, Coeficiente de variación.

A. Rango de variación

Se define como la diferencia entre el mayor valor de la variable y el menor valor de la variable.

Rango de Variación = Máximo - Mínimo

B. Varianza y desviación estándar

La mejor medida de dispersión, y la más generalizada es la varianza, o su raíz cuadrada, la desviación estándar. La varianza se representa con el símbolo σ^2 (sigma cuadrado) para el universo o población y con el símbolo s2 (s cuadrado), cuando se trata de la muestra. La desviación estándar, que es la raíz cuadrada de la varianza, se representa por σ (sigma) cuando pertenece al universo o población y por "s", cuando pertenece a la muestra. σ^2 y σ son parámetros, constantes para una población particular; s2 y s son estadígrafos, valores que cambian de muestra en muestra dentro de una misma población.

Fórmulas
Donde μ es el promedio de la población.

$$\sigma^2 = \frac{\sum_i (Y_i - u)^2}{N}$$

Donde $\bar{Y}$ es el promedio de la muestra.

$$s^2 = \frac{\sum_i (Y_i - \bar{Y})^2}{n - 1}$$

Consideremos a modo de ejemplo una muestra de 4 observaciones

Según la fórmula el promedio calculado es 7, veamos ahora el cálculo de las medidas de dispersión:

Y_i	$Y_i - \bar{Y}$	$(Y_i - \bar{Y})^2$
3	-4	16
6	-1	1
8	+1	1
11	+4	16
Suma		34

- ✓ S^2 = 34 / 3 = 11,33 **Varianza de la muestra**
- ✓ La **desviación estándar** de la muestra (s) será la raíz cuadrada de 11,33 = 3,4.

Interpretación de la varianza (válida también para la desviación estándar): un alto valor de la varianza indica que los datos están alejados del promedio. Es difícil hacer una interpretación de la varianza teniendo un solo valor de ella. La situación es más clara si se comparan las varianzas de dos muestras, por ejemplo, varianza de la muestra igual 18 y varianza de la muestra b igual 25. En este caso diremos que los datos de la muestra b tienen mayor dispersión que los datos de la muestra a. esto significa que en la muestra a los datos están más cerca del promedio y en cambio en la muestra b los datos están más alejados del promedio.

C. Coeficiente de variación

Es una medida de la dispersión relativa de los datos. Se define como la desviación estándar de la muestra expresada como porcentaje de la media muestral.

$$CV = \frac{\sigma}{\mu} \cdot 100\%$$

Donde:
CV = Coeficiente de variación.
σ = desviación estándar de la población.
μ = media aritmética de la población.

Es de particular utilidad para comparar la dispersión entre variables con distintas unidades de medida. Esto porque el coeficiente de variación, a diferencia de la desviación estándar, es independiente de la unidad de medida de la variable de estudio.

1.2.4. Medidas de tendencia central y de dispersión en datos agrupados

Se identifica como datos agrupados a los datos dispuestos en una distribución de frecuencia. En tal caso las fórmulas para el cálculo de promedio, mediana, modo, varianza y desviación estándar deben incluir una leve modificación. A continuación, se entregan los detalles para cada una de las medidas.

A. Promedio en datos agrupados

La fórmula es la siguiente:

$$\overline{Y} = \frac{\sum_{i=1}^{n} Y_i n_i}{n}$$

- ✓ Donde ni representa cada una de las frecuencias correspondientes a los diferentes valores de Yi.

Ejemplo:

Una distribución de frecuencia de madres que asisten a un programa de lactancia materna, clasificadas según el número de partos. Por tratarse de una variable en escala discreta, las clases o categorías asumen sólo ciertos valores: 1, 2, 3, 4, 5.

Yi- nº de partos	ni	Yi ni	Ni (Frec acumulada)
1	4	4	4
2	13	26	17
3	16	48	33
4	6	24	39
5	3	15	42
Total	42	117	

$$\overline{Y} = \frac{\sum_{i=1}^{n} Y_i n_i}{n} = \frac{117}{42} = 2{,}78$$

Entonces las 42 madres han tenido, en promedio, 2,78 partos.

Si la variable de interés es de tipo continuo será necesario determinar, para cada intervalo, un valor medio que lo represente. Este valor se llama marca de clase (Yc) y se calcula dividiendo por 2 la suma de los límites reales del intervalo de clase. De ahí en adelante se procede del mismo modo que en el ejercicio anterior, reemplazando, en la fórmula de promedio, Yi por Yc.

B. Mediana en datos agrupados

Si la variable es de tipo discreto la mediana será el valor de la variable que corresponda a la frecuencia acumulada que supere inmediatamente a n/2. En los datos de la tabla 1 Me=3, ya que 42/2 es igual a 21 y la frecuencia acumulada que supera inmediatamente a 21 es 33, que corresponde a un valor de variable (Yi) igual a 3.

Si la variable es de tipo continuo es necesario, primero, identificar la frecuencia acumulada que supere en forma inmediata a n/2, y luego aplicar la siguiente fórmula:

$$Me = L_{i-1} + \frac{\frac{N}{2} - F_{i-1}}{fi} \cdot a$$

Donde:

L_{i-1} = *Límite inferior del intervalo mediana*
a = *Amplitud del intervalo mediana*
F_{i-1} = *Frecuencia acumulada anterior al intervalo mediana*
f_i = *Frecuencia absoluta del intervalo mediana*
N = *Total de datos*

C. Moda en datos agrupados

Si la variable es de tipo discreto la moda o modo será al valor de la variable (Yi) que tenga la mayor frecuencia absoluta . En los datos de la tabla 1 el valor de la moda es 3 ya que este valor de variable corresponde a la mayor frecuencia absoluta =16.

Más adelante se presenta un ejemplo integrado para promedio, mediana, varianza y desviación estándar en datos agrupados con intervalos.

D. Varianza en datos agrupados

Para el cálculo de varianza en datos agrupados se utiliza la fórmula

$$s^2 = \frac{\sum_i (y_i - \bar{y})^2 n_i}{n-1}$$

Con los datos del ejemplo y recordando que el promedio (Y) resultó ser 2,78 partos por madre.

Y_i	n_i	$Y_i\, n_i$	$(Y_i - \bar{Y})^2$	$(y_i - \bar{y})^2 n_i$
1	4	4	3,1684	12,67
2	13	26	0,6084	7,9
3	16	48	0,0484	0,7744
4	6	24	1,4884	8,93
5	3	15	4,9284	14,7852
Total	42	117		45,06

$$s^2 = \frac{\sum_i (y_i - \bar{y})^2 n_i}{n-1} = 45{,}06 \,/\, 42\text{-}1 = 45{,}06 \,/41 = 1{,}1$$

Cuando los datos están agrupados en intervalos de clase, se trabaja con la marca de clase (Yc), de tal modo que la fórmula queda:

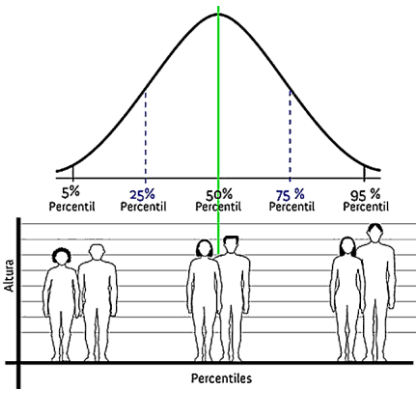
$$s^2 = \frac{\sum_i (y_c - \bar{y})^2 n_i}{n-1}$$

Donde Yc es el punto medio del intervalo y se llama marca de clase del intervalo

Yc= (Límite inferior del intervalo + límite superior del intervalo) /2.

E. Percentiles

Los percentiles son valores de la variable que dividen la distribución en 100 partes iguales. De este modo si el percentil 80 (P80) es igual a 35 años, significa que el 80% de los casos tiene edad igual o inferior a 35 años.



Su procedimiento de cálculo es relativamente simple en datos agrupados sin intervalos.

Retomemos el ejemplo de la variable número de partos:

Yi- nº de partos	ni	Ni
1	4	4
2	13	17
3	16	33
4	6	39
5	3	42
Total	42	

El percentil j (Pj) corresponde al valor de la variable (Yi) cuya frecuencia acumulada supera inmediatamente al "j" % de los casos (jxn/100).

El percentil 80, en los datos de la tabla, será el valor de la variable cuyo Ni sea inmediatamente superior a 33,6 ((80x42) /100).

El primer Ni que supera a 33,6 es 39. Por lo tanto, al percentil 80 le corresponde el valor 4. Se dice entonces que el percentil 80 es 4 partos (P80=4). Este resultado significa que un 80% de las madres estudiadas han tenido 4 partos o menos.

Si los datos están agrupados en una tabla con intervalos, el procedimiento es levemente más complejo ya que se hace necesaria la aplicación de una fórmula.

$$P_j = LI + \frac{\left[\frac{jn}{100} - N_{i-1}\right]}{n_i} A_i$$

Se aplica a los datos del intervalo cuya frecuencia acumulada (Ni) sea inmediatamente superior al "j" % de los casos (jxn/100).

En la siguiente tabla se muestra la distribución de 40 familias según su ingreso mensual en miles de pesos. Nótese que para calcular el centro de clase se usaron los límites reales de cada intervalo.

Ingreso Familiar (miles)	n_i	Y_c	$y_c \times n_i$	$(y_c - \bar{y})$	$(y_c - \bar{y})^2$	$(y_c - \bar{y})^2 \times n_i$	A_i	N_i
80 - 99	8	90	720	-49,13	2.413,76	19.310,08	20	8
100 - 119	10	110	1100	-29,13	848,56	8.485,60	20	18
120 - 159	11	140	1540	0,87	0,76	8,36	40	29
160 - 199	6	180	1080	40,87	1.670,36	10.022,16	40	35
200 - 249	5	225	1125	85,87	7.373,66	36.868,30	50	40
Total	40		5565			74.694,50		

El ingreso mensual promedio será:

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_c n_i}{n} = \frac{5565}{40} = 139,13$$

La mediana será:

$$Me = LI + \frac{\left[\frac{n}{2} - N_{i-1}\right]}{n_i} A_i = 120 + \frac{\left[\frac{40}{2} - 18\right]}{11} \times 40 = 120 + 7,27 = 127,27$$

Esto significa que un 50% de las familias tiene ingreso mensual igual o inferior a $127.270.

El percentil 78 será:

$$P_{78} = LI + \frac{\left[\frac{78n}{100} - N_{i-1}\right]}{n_i} A_i = 160 + \frac{[31{,}2 - 29]}{6} 40 = 160 + 14{,}66 = 174{,}66$$

Por lo tanto, se puede decir que 78% de las familias tienen ingreso igual o inferior a $174.660.

Los percentiles 10 y 90 serán:

$$P10 = LI + \frac{\left[\frac{10n}{100} - N_{i-1}\right]}{n_i} A_i = 80 + \frac{[4 - 0]}{8} 20 = 80 + 10 = 90$$

$$P90 = LI + \frac{\left[\frac{90n}{100} - N_{i-1}\right]}{n_i} A_i = 200 + \frac{[36 - 35]}{5} 50 = 200 + 10 = 210$$

A base de los valores de los percentiles 10 y 90 se pueden hacer tres afirmaciones:

- ✓ El 10% de las familias tiene ingreso igual o inferior a $90.000.
- ✓ El 90% de las familias tiene ingreso igual o inferior a $210.000.
- ✓ El 80% central, de las familias, tiene ingreso entre $90.000 y $210.000

La varianza será:

$$s^2 = \frac{\sum_i (y_i - \bar{y})^2 n_i}{n-1} = 74.694{,}5 / 39 = 1915{,}24$$

La desviación estándar es la raíz cuadrada de esta cifra, es decir: 43,76.

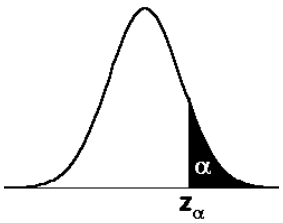
1.2.5. Distribución muestral de medias

Si tenemos una muestra aleatoria de una población N(μ,σ), se sabe (Teorema del límite central) que la fdp de la media muestral es también normal con media μ y varianza σ^2/n. Esto es exacto para poblaciones normales y aproximado (buena aproximación con n>30) para poblaciones cualesquiera. Es decir $\sigma/\sqrt{n}$ es el error típico, o error estándar de la media.

¿Cómo usamos esto en nuestro problema de estimación? 1º problema: No hay tablas para cualquier normal, sólo para la normal μ=0 y σ=1 (la llamada z); pero haciendo la transformación (llamada tipificación)

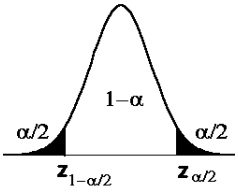
$z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$ una normal de media μ y desviación σ se transforma en una z.

Llamando z_α al valor de una variable normal tipificada que deja a su derecha un área bajo la curva de α, es decir, que la probabilidad que la variable sea mayor que ese valor es α(estos son los valores que ofrece la tabla de la normal)



podremos construir intervalos de la forma

$$z_{1-\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}$$



para los que la probabilidad es 1 - α.

Teniendo en cuenta la simetría de la normal y manipulando algebraicamente

$$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

que también se puede escribir

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

o, haciendo énfasis en que $\sigma/\sqrt{n}$ es el error estándar de la media,

$$\bar{X} \pm z_{\alpha/2} EE(\bar{X})$$

Recuérdese que la probabilidad de que μ esté en este intervalo es $1 - \alpha$. A un intervalo de este tipo se le denomina intervalo de confianza con un nivel de confianza del $100(1 - \alpha)\%$, o nivel de significación de $100\alpha\%$. El nivel de confianza habitual es el 95%, en cuyo caso $\alpha=0{,}05$ y $z_{\alpha/2}=1{,}96$. Al valor $\bar{X}$ se le denomina estimación puntual y se dice que $\bar{X}$ es un estimador de μ.

EJEMPLO: Si de una población normal con varianza 4 se extrae una muestra aleatoria de tamaño 20 en la que se calcula $\bar{X} = 5{,}3$ se puede decir que μtiene una probabilidad de 0,95 de estar comprendida en el intervalo

$$5{,}3 \pm 1{,}96\frac{2}{\sqrt{20}} = (4{,}42 \quad 6{,}18)$$

que sería el intervalo de confianza al 95% para μ

En general esto es poco útil, en los casos en que no se conoce μ tampoco suele conocerse σ^2; en el caso más realista de σ^2 desconocida los intervalos de confianza se construyen con la *t de Student* (otra *fdp*continua para la que hay tablas) en lugar de la *z*.

$$\bar{X} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$$

o, haciendo énfasis en que $s/\sqrt{n}$ es el error estándar estimado de la media,

$$\bar{X} \pm t_{\alpha/2} E\hat{E}\left(\bar{X}\right)$$

Esta manera de construir los intervalos de confianza sólo es válida si la variable es normal. Cuando *n* es grande (>30) se puede sustituir *t* por *z* sin mucho error.

1.3. Datos discretos y continuos

Al trabajar con estadísticas, es importante reconocer los diferentes tipos de datos: numéricos (discretos y continuos), categóricos y ordinales.

Los datos son las piezas de información reales que recopila a través de su estudio. Por ejemplo, si le preguntas a cinco de tus amigos cuántas mascotas tienen, pueden darte los siguientes datos: 0, 2, 1, 4, 18.

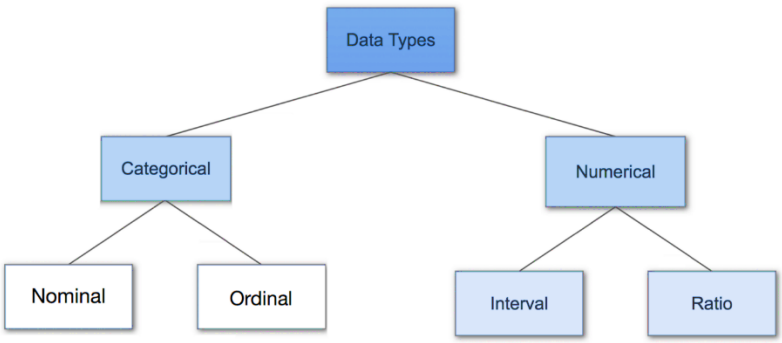
No todos los datos son números, digamos que también registras el sexo de cada uno de tus amigos, obteniendo los siguientes datos: masculino, masculino, femenino, masculino, femenino.

La mayoría de los datos caen en uno de dos grupos: **numéricos o categóricos**.

Pero aún es importante tener al menos una comprensión básica de los diferentes tipos de datos y los tipos de preguntas que puede usar para responder.

A continuación, se explica cómo podemos recopilar estos diferentes tipos de datos, y para qué tipo de análisis podríamos usarlo.

1.3.1. Los dos principales tipos de datos: cualitativos y cuantitativos



En el nivel más alto, existen dos tipos de datos: cuantitativo y cualitativo.

La información cuantitativa trata de números y cosas que puede medir objetivamente: dimensiones como altura, ancho y longitud. Temperatura y humedad. Precios. Área y volumen

Los datos cualitativos tratan con características y descripciones que no se pueden medir fácilmente, pero que se pueden observar subjetivamente, como los olores, los sabores, las texturas, el atractivo y el color.

En términos generales, cuando mides algo y le das un valor numérico, creas datos cuantitativos. Cuando clasifica o juzga algo, crea datos cualitativos. Hasta aquí todo bien. Pero este es solo el nivel más alto de datos: también hay diferentes tipos de datos cuantitativos y cualitativos.

1.3.2. Tipos cuantitativos: datos continuos y datos discretos

Hay dos tipos de datos cuantitativos, que también se conocen como datos numéricos: continuo y discreto. Como regla general, los recuentos son discretos y las mediciones son continuas.

Los datos discretos son un conteo que no se puede hacer más preciso. Por lo general, implica números enteros. Por ejemplo, el número de niños (o adultos, o mascotas) en su familia es información discreta, porque está contando entidades enteras e indivisibles: no puede tener 2.5 hijos o 1.3 mascotas.

Los datos continuos, por otro lado, podrían dividirse y reducirse a niveles cada vez más finos. Por ejemplo, puede medir la altura de sus hijos en escalas progresivamente más precisas (metros, centímetros, milímetros y más), por lo que la altura es un dato continuo.

Si cuento el número de manís individuales en una caja, ese número es una pieza de datos discretos.

Si utilizo una balanza para medir el peso de cada maní, o el peso de toda la caja, eso es información continua.

Los datos continuos se pueden usar en muchos tipos diferentes de pruebas de hipótesis. Por ejemplo, para evaluar la precisión del peso impreso en la caja de manís, podríamos medir 30 cajas y realizar una prueba t de 1 muestra.

Algunos análisis usan datos cuantitativos continuos y discretos al mismo tiempo.

Por ejemplo, podríamos realizar un análisis de regresión para ver si el peso de los manís (datos continuos) está correlacionado con el número de manís en el interior (datos discretos).

1.3.3. Tipos Cualitativos: Datos Binomiales, Datos Nominales y Datos Ordinales

Cuando se clasifica o categoriza algo, se crea datos cualitativos o de atributos. Hay tres tipos principales de datos cualitativos.

Los datos binarios colocan las cosas en una de dos categorías mutuamente excluyentes: correcto / incorrecto, verdadero / falso o aceptar / rechazar.

Ocasionalmente, obtendrémos una caja de manís que contiene un par de piezas individuales que son demasiado duras o demasiado secas. Si revisé la casilla y clasifiqué cada pieza como "Bueno" o "Malo", serían datos binarios.

Podría usar este tipo de datos para desarrollar un modelo estadístico para predecir con qué frecuencia puedo esperar obtener un maní malo.

Al recopilar datos no ordenados o nominales, asignamos elementos individuales a categorías con nombre que no tienen un valor o rango implícito o natural. Si revisé una caja de manís y grabé el color de cada una en mi hoja de trabajo, serían datos nominales.

Este tipo de datos se puede usar de muchas maneras diferentes; por ejemplo, podría usar el análisis de chi-cuadrado para ver si hay diferencias estadísticamente significativas en las cantidades de cada color en un cuadro.

También podemos tener datos ordenados u ordinales, en los que los elementos se asignan a categorías que tienen algún tipo de orden implícito o natural, como "Corto, Medio o Alto".

Otro ejemplo es una pregunta de la encuesta que nos pide que califiquemos un artículo en una escala de 1 a 10, siendo 10 el mejor. Esto implica que 10 es mejor que 9, que es mejor que 8, y así sucesivamente.

Los usos de los datos ordenados son un tema de debate entre los estadísticos.

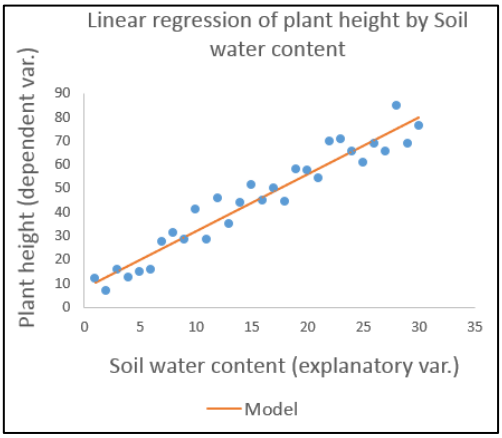
Todos coinciden en que es apropiado para crear gráficos de barras, pero más allá de eso, no se sabe más.

1.3.4. Modelado estadístico, inspección científica y representación grafica de datos

En términos sencillos, el **modelado estadístico** es una forma simplificada, matemáticamente formalizada, de aproximarse a la realidad (i.e., la que genera los datos) y, opcionalmente, hacer predicciones a partir de dicha aproximación.

Veamos un ejemplo básico. Supongamos que deseamos informar sobre el peso de una variedad de patatas. Podemos considerar una forma difícil y otra fácil de hacerlo. La difícil consiste en emplear años midiendo el peso de cada patata de esta variedad a lo largo del mungo, y trasladar los datos a una hoja Excel interminable. La fácil, consiste en seleccionar 30 patatas ampliamente representativas de la variedad que nos interesa, calcular su media y su desviación estándar, e informar únicamente sobre esas magnitudes como una descripción aproximada del peso. Representar una cantidad por una media y una desviación estándar es una forma muy simple de modelado estadístico.

Otro ejemplo es intentar representar la altura de las plantas en función del agua del suelo mediante una línea recta caracterizada por una pendiente y una intercepción, trazada después de realizar un experimento sobre una muestra de plantas sometidas a una humedad del suelo creciente. Este modelo particular se denomina regresión lineal simple.



En la mayoría de los casos, los modelos estadísticos implican variables explicativas y variables dependientes.

La **variable dependiente** es aquella que queremos describir, explicar, o predecir. Como norma general, la variable dependiente es la representada en el eje Y en los gráficos. En el ejemplo de la altura de las plantas, la variable dependiente es la altura de la planta. Las **variables explicativas**, también denominadas variables **independientes**, son aquellas que usamos para explicar, describir o predecir la(s) variable(s) dependiente(s). Las variables explicativas generalmente se representan en el eje X. En el ejemplo de la altura de las plantas solo usamos una variable independiente cuantitativa (el contenido en agua del suelo). Tanto las variables dependientes como las explicativas pueden ser una o varias, cuantitativas o cualitativas. Se han desarrollado modelos adaptados a las diferentes situaciones.

A. ¿Qué es un parámetro de un modelo?

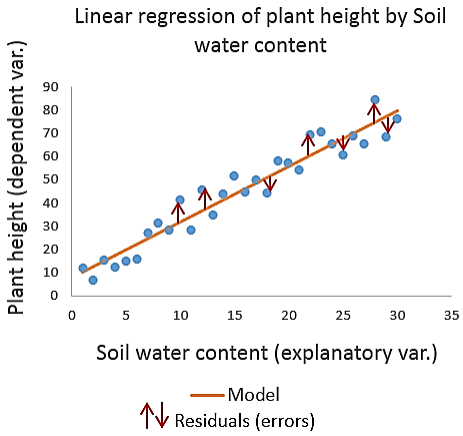
En los modelos clásicos, paramétricos, la(s) variable(s) dependiente(s) está vinculada a las explicativas a través de una ecuación matemática (el modelo) que implica cantidades denominadas **parámetros del modelo**. En el ejemplo de la altura de las plantas (regresión lineal simple) los parámetros son la intercepción y la pendiente. La ecuación puede representarse así:

Altura = **intercepción** + **pendiente***contenido en agua del suelo

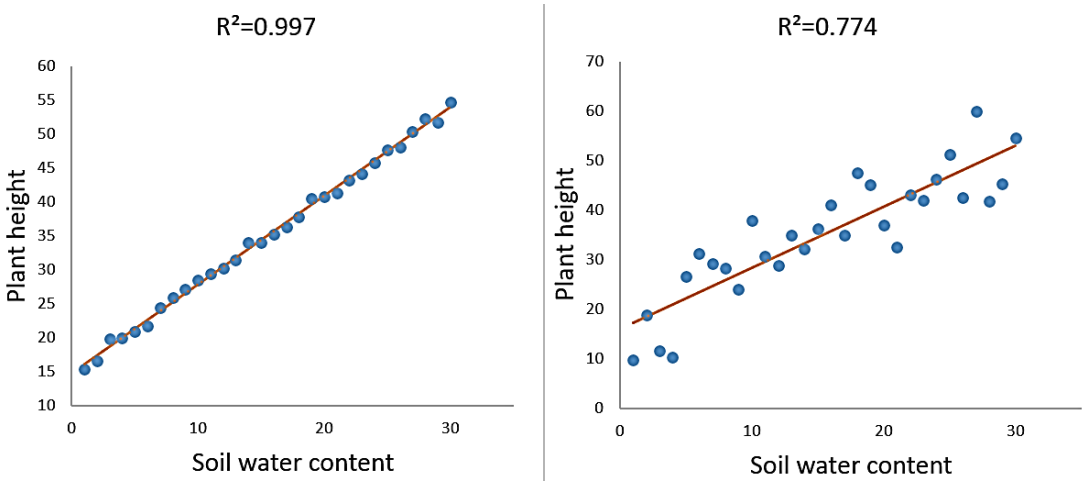
Los cálculos que subyacen al modelado estadístico permiten la estimación de los parámetros del modelo, así como ulteriores predicciones de la variable dependiente. [1]La regresión lineal simple implica también un tercer parámetro, la varianza de los residuos (vid. párrafo siguiente).

B. ¿Qué es un residuo de un modelo?

Técnicamente, los **residuos** (o errores) de un modelo son las distancias entre los puntos que representan los datos y el modelo (que está representado por una línea recta en el ejemplo de regresión lineal simple de la altura de las plantas).



Los residuos del modelo representan la parte de variabilidad de los datos que el modelo ha sido incapaz de capturar. El estadístico R^2 es la parte de variabilidad que es explicada por el modelo. Así, mientras más pequeños sean los residuos, mayor será el estadístico R^2.



C. ¿Qué modelo estadístico debería elegir?

Esta tabla le guiará en la elección de los modelos más frecuentemente usados de acuerdo al tipo y número de variables dependientes e independientes. También se proponen soluciones distintas a los modelos paramétricos.

D. Representación gráfica en el Análisis de Datos

La realización de los estudios clínico-epidemiológicos implica finalmente emitir unos resultados cuantificables de dicho estudio o experimento. La claridad de dicha presentación es de vital importancia para la comprensión de los resultados y la interpretación de los mismos. A la hora de representar los resultados de un análisis estadístico de un modo adecuado, son varias las publicaciones que podemos consultar1. Aunque se aconseja que la presentación de datos numéricos se haga habitualmente por medio de tablas, en ocasiones un diagrama o un gráfico pueden ayudarnos a representar de un modo más eficiente nuestros datos.

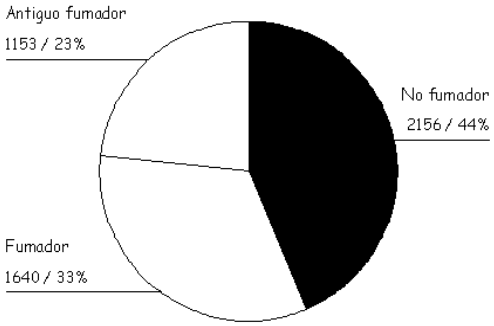
Se abordará la representación gráfica de los resultados de un estudio, constatando su utilidad en el proceso de análisis estadístico y la presentación de datos. Se describirán los distintos tipos de gráficos que podemos utilizar y su correspondencia con las distintas etapas del proceso de análisis.

- ✓ **Análisis descriptivo.**

Cuando se dispone de datos de una población, y antes de abordar análisis estadísticos más complejos, un primer paso consiste en presentar esa información de forma que ésta se pueda visualizar de una manera más sistemática y resumida. Los datos que nos interesan dependen, en cada caso, del tipo de variables que estemos manejando2.

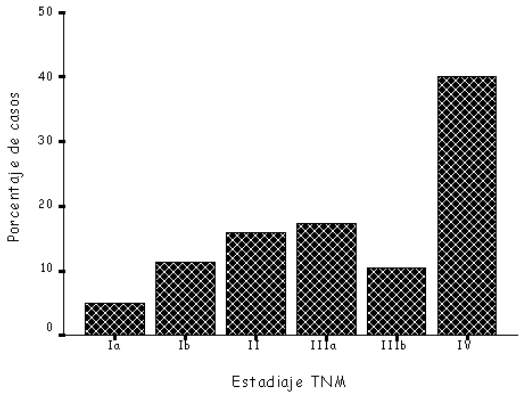
Para variables categóricas, como el sexo, profesión, etc., se quiere conocer la frecuencia y el porcentaje del total de casos que "caen" en cada categoría. Una forma muy sencilla de representar gráficamente estos resultados es mediante diagramas de barras o diagramas de sectores. En los gráficos de sectores, también conocidos como diagramas de "tartas", se divide un círculo en tantas porciones como clases tenga la variable, de modo que a cada clase le corresponde un arco de círculo proporcional a su frecuencia absoluta o relativa. Un ejemplo se muestra en la Figura. Como se puede observar, la información que se debe mostrar en cada sector hace referencia al número de casos dentro de cada categoría y al porcentaje del total que estos representan. Si el número de categorías es excesivamente grande, la imagen proporcionada por el gráfico de sectores no es lo suficientemente clara y por lo tanto la situación ideal es cuando hay alrededor de tres categorías. En este caso se pueden apreciar con claridad dichos subgrupos.

Ejemplo de gráfico de sectores. Distribución de una muestra de pacientes según el hábito de fumar.



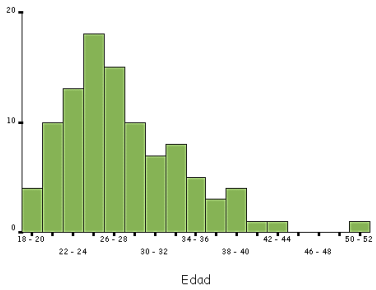
Los diagramas de barras son similares a los gráficos de sectores. Se representan tantas barras como categorías tiene la variable, de modo que la altura de cada una de ellas sea proporcional a la frecuencia o porcentaje de casos en cada clase . Estos mismos gráficos pueden utilizarse también para describir variables numéricas discretas que toman pocos valores (número de hijos, número de recidivas, etc.).

Ejemplo de gráfico de barras. Estadio TNM en el cáncer gástrico.



Para variables numéricas continuas, tales como la edad, la tensión arterial o el índice de masa corporal, el tipo de gráfico más utilizado es el histograma. Para construir un gráfico de este tipo, se divide el rango de valores de la variable en intervalos de igual amplitud, representando sobre cada intervalo un rectángulo que tiene a este segmento como base. El criterio para calcular la altura de cada rectángulo es el de mantener la proporcionalidad entre las frecuencias absolutas (o relativas) de los datos en cada intervalo y el área de los rectángulos. Como ejemplo, la Tabla Imuestra la distribución de frecuencias de la edad de 100 pacientes, comprendida entre los 18 y 42 años. Si se divide este rango en intervalos de dos años, el primer tramo está comprendido entre los 18 y 19 años, entre los que se encuentra el 4/100=4% del total. Por lo tanto, la primera barra tendrá altura proporcional a 4. Procediendo así sucesivamente, se construye el histograma que se muestra en la figura. Uniendo los puntos medios del extremo superior de las barras del histograma, se obtiene una imagen que se llama polígono de frecuencias. Dicha figura pretende mostrar, de la forma más simple, en qué rangos se encuentra la mayor parte de los datos. Un ejemplo, utilizando los datos anteriores, se presenta en la figura

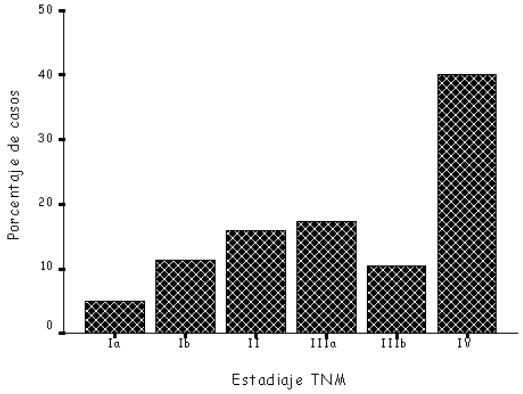
Ejemplo de un histograma correspondiente a los datos de la Tabla



Polígono de frecuencias para los datos de la Tabla

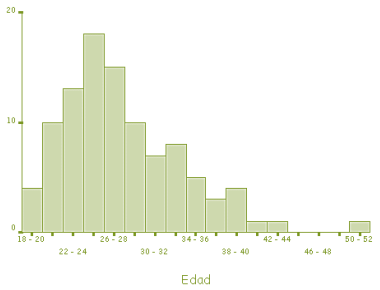
Los diagramas de barras son similares a los gráficos de sectores. Se representan tantas barras como categorías tiene la variable, de modo que la altura de cada una de ellas sea proporcional a la frecuencia o porcentaje de casos en cada clase . Estos mismos gráficos pueden utilizarse también para describir variables numéricas discretas que toman pocos valores (número de hijos, número de recidivas, etc.).

Ejemplo de gráfico de barras. Estadio TNM en el cáncer gástrico.

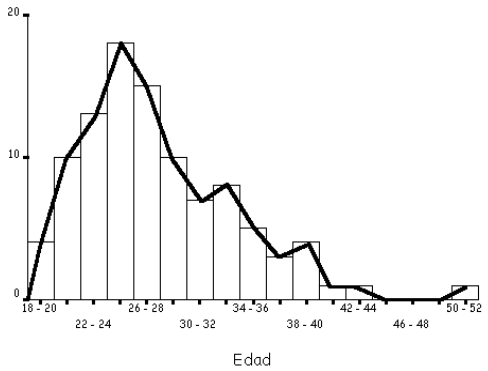


Para variables numéricas continuas, tales como la edad, la tensión arterial o el índice de masa corporal, el tipo de gráfico más utilizado es el histograma. Para construir un gráfico de este tipo, se divide el rango de valores de la variable en intervalos de igual amplitud, representando sobre cada intervalo un rectángulo que tiene a este segmento como base. El criterio para calcular la altura de cada rectángulo es el de mantener la proporcionalidad entre las frecuencias absolutas (o relativas) de los datos en cada intervalo y el área de los rectángulos. Como ejemplo, la Tabla Imuestra la distribución de frecuencias de la edad de 100 pacientes, comprendida entre los 18 y 42 años. Si se divide este rango en intervalos de dos años, el primer tramo está comprendido entre los 18 y 19 años, entre los que se encuentra el 4/100=4% del total. Por lo tanto, la primera barra tendrá altura proporcional a 4. Procediendo así sucesivamente, se construye el histograma que se muestra en la Figura . Uniendo los puntos medios del extremo superior de las barras del histograma, se obtiene una imagen que se llama polígono de frecuencias. Dicha figura pretende mostrar, de la forma más simple, en qué rangos se encuentra la mayor parte de los datos. Un ejemplo, utilizando los datos anteriores, se presenta en la Figura .

Ejemplo de un histograma correspondiente a los datos de la Tabla

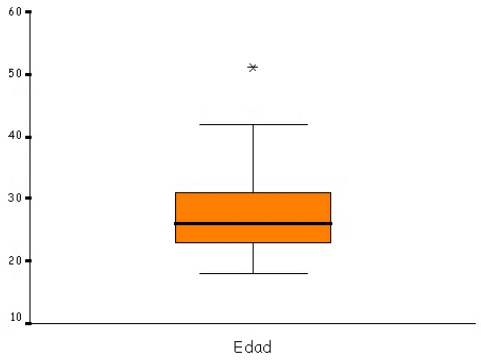


Polígono de frecuencias para los datos de la Tabla I.



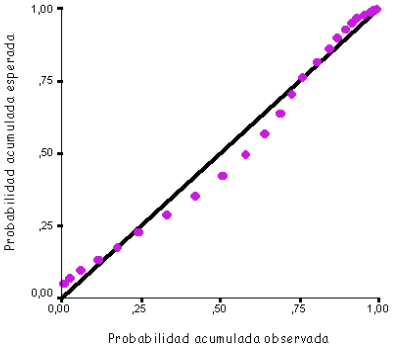
Otro modo habitual, y muy útil, de resumir una variable de tipo numérico es utilizando el concepto de percentiles, mediante diagramas de cajas4,5. La Figura muestra un gráfico de cajas correspondiente a los datos de la Tabla I. La caja central indica el rango en el que se concentra el 50% central de los datos. Sus extremos son, por lo tanto, el 1er y 3er cuartil de la distribución. La línea central en la caja es la mediana. De este modo, si la variable es simétrica, dicha línea se encontrará en el centro de la caja. Los extremos de los "bigotes" que salen de la caja son los valores que delimitan el 95% central de los datos, aunque en ocasiones coinciden con los valores extremos de la distribución. Se suelen también representar aquellas observaciones que caen fuera de este rango (outliers o valores extremos). Esto resulta especialmente útil para comprobar, gráficamente, posibles errores en nuestros datos. En general, los diagramas de cajas resultan más apropiados para representar variables que presenten una gran desviación de la distribución normal. Como se verá más adelante, resultan además de gran ayuda cuando se dispone de datos en distintos grupos de sujetos.

Ejemplo de un diagrama de caja correspondiente a lo datos en la Tabla I.



Por último, y en lo que respecta a la descripción de los datos, suele ser necesario, para posteriores análisis, comprobar la normalidad de alguna de las variables numéricas de las que se dispone. Un diagrama de cajas o un histograma son gráficos sencillos que permiten comprobar, de un modo puramente visual, la simetría y el "apuntamiento" de la distribución de una variable y, por lo tanto, valorar su desviación de la normalidad. Existen otros métodos gráficos específicos para este propósito, como son los gráficos P-P o Q-Q. En los primeros, se confrontan las proporciones acumuladas de una variable con las de una distribución normal. Si la variable seleccionada coincide con la distribución de prueba, los puntos se concentran en torno a una línea recta. Los gráficos Q-Q se obtienen de modo análogo, esta vez representando los cuantiles de distribución de la variable respecto a los cuantiles de la distribución normal. En la Figura se muestra el gráfico P-P correspondientes a los datos de la Tabla I que sugiere, al igual que el correspondiente histograma y el diagrama de cajas, que la distribución de la variable se aleja de la normalidad.

Gráfico P-P de normalidad para los datos de la Tabla I.

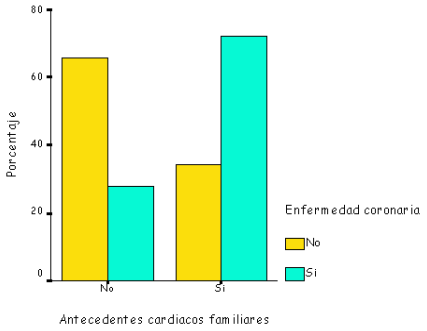


Comparación de dos o más grupos.

Cuando se quieren comparar las observaciones tomadas en dos o más grupos de individuos una vez más el método estadístico a utilizar, así como los gráficos apropiados para visualizar esa relación, dependen del tipo de variables que estemos manejando.

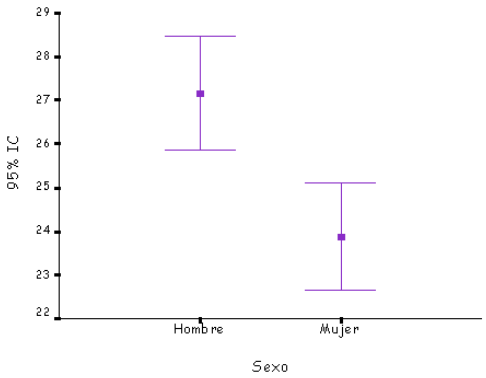
Cuando se trabaja con dos variables cualitativas podemos seguir empleando gráficos de barras o de sectores. Podemos querer determinar, por ejemplo, si en una muestra dada, la frecuencia de sujetos que padecen una enfermedad coronaria es más frecuente en aquellos que tienen algún familiar con antecedentes cardiacos. A partir de dicha muestra podemos representar, como se hace en la Figura , dos grupos de barras: uno para los sujetos con antecedentes cardiacos familiares y otro para los que no tienen este tipo de antecedentes. En cada grupo, se dibujan dos barras representando el porcentaje de pacientes que tienen o no alguna enfermedad coronaria. No se debe olvidar que cuando los tamaños de las dos poblaciones son diferentes, es conveniente utilizar las frecuencias relativas, ya que en otro caso el gráfico podría resultar engañoso.

Diagrama de barras agrupadas. Relación entre la presencia de alguna enfermedad coronaria y los antecedentes cardiacos familiares en una muestra.



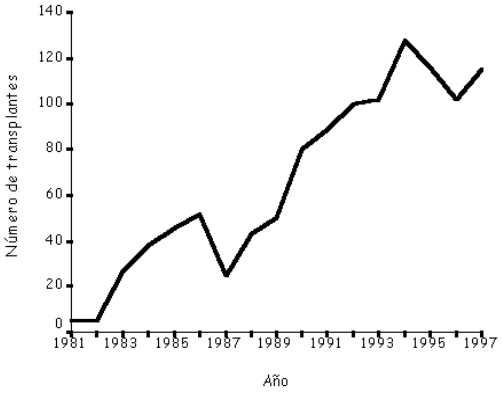
Por otro lado, la comparación de variables continuas en dos o más grupos se realiza habitualmente en términos de su valor medio, por medio del test t de Student, análisis de la varianza o métodos no paramétricos equivalentes, y así se ha de reflejar en el tipo de gráfico utilizado. En este caso resulta muy útil un diagrama de barras de error, como en la Figura . En él se compara el índice de masa corporal en una muestra de hombres y mujeres. Para cada grupo, se representa su valor medio, junto con su 95% intervalo de confianza. Conviene recordar que el hecho de que dichos intervalos no se solapen, no implica necesariamente que la diferencia entre ambos grupos pueda ser estadísticamente significativa, pero sí nos puede servir para valorar la magnitud de la misma. Así mismo, para visualizar este tipo de asociaciones, pueden utilizarse dos diagramas de cajas, uno para cada grupo. Estos diagramas son especialmente útiles aquí: no sólo permiten ver si existe o no diferencia entre los grupos, sino que además nos permiten comprobar la normalidad y la variabilidad de cada una de las distribuciones. No olvidemos que las hipótesis de normalidad y homocedasticidad son condiciones necesarias para aplicar algunos de los procedimientos de análisis paramétricos.

Barras de error. Variación en el índice de masa corporal según el sexo.



Por último, señalar que también en esta situación pueden utilizarse los ya conocidos gráficos de barras, representando aquí como altura de cada barra el valor medio de la variable de interés. Los gráficos de líneas pueden resultar también especialmente interesantes, sobre todo cuando interesa estudiar tendencias a lo largo del tiempo (Figura). No son más que una serie de puntos conectados entre sí mediante rectas, donde cada punto puede representar distintas cosas según lo que nos interese en cada momento (el valor medio de una variable, porcentaje de casos en una categoría, el valor máximo en cada grupo, etc).

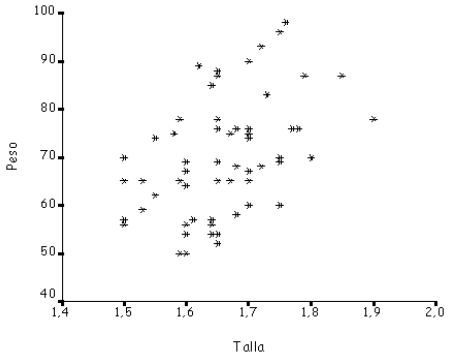
Gráfico de líneas. Número de pacientes trasplantados renales en el Complexo Hospitalario "Juan Canalejo" durante el periodo 1981-1997.



Relación entre dos variables numéricas.

Cuando lo que interesa es estudiar la relación entre dos variables continuas, el método de análisis adecuado es el estudio de la correlación. Los coeficientes de correlación (Pearson, Spearman, etc.) valoran hasta qué punto el valor de una de las variables aumenta o disminuye cuando crece el valor de la otra. Cuando se dispone de todos los datos, un modo sencillo de comprobar, gráficamente, si existe una correlación alta, es mediante diagramas de dispersión, donde se confronta, en el eje horizontal, el valor de una variable y en el eje vertical el valor de la otra. Un ejemplo sencillo de variables altamente correlacionados es la relación entre el peso y la talla de un sujeto. Partiendo de una muestra arbitraria, podemos construir el diagrama de dispersión de la Figura 10. En él puede observarse claramente como existe una relación directa entre ambas variables, y valorar hasta qué punto dicha relación puede modelizarse por la ecuación de una recta. Este tipo de gráficos son, por lo tanto, especialmente útiles en la etapa de selección de variables cuando se ajusta un modelo de regresión lineal.

Figura 10. Diagrama de dispersión entre la talla y el peso de una muestra de individuos.



Otros gráficos.

Los tipos de gráficos mostrados hasta aquí son los más sencillos que podemos manejar, pero ofrecen grandes posibilidades para la representación de datos y pueden ser utilizados en múltiples situaciones, incluso para representar los resultados obtenidos por métodos de análisis más complicados. Podemos utilizar, por ejemplo, dos diagramas de líneas superpuestos para visualizar los resultados de un análisis de la varianza con dos factores (Figura 11). Un diagrama de dispersión es el método adecuado para valorar el resultado de un modelo de regresión logística (Figura 12). Existen incluso algunos análisis concretos que están basados completamente en la representación gráfica. En particular, la elaboración de curvas ROC (Figura 13) y el cálculo del área bajo la curva constituyen el método más apropiado para valorar la exactitud de una prueba diagnóstica.

Hemos visto, por lo tanto, como la importancia y utilidad que las representaciones gráficas pueden alcanzar en el proceso de análisis de datos. La mayoría de los textos estadísticos y epidemiológicos4 hacen hincapié en los distintos tipos de gráficos que se pueden crear, como una herramienta imprescindible en la presentación de resultados y el proceso de análisis

estadístico. No obstante, es difícil precisar cuándo es más apropiado utilizar un gráfico que una tabla. Más bien podremos considerarlos dos modos distintos pero complementarios de visualizar los mismos datos. La creciente utilización de distintos programas informáticos hace especialmente sencillo la obtención de las mismas. La mayoría de los paquetes estadísticos (SPSS, STATGRAPHICS, S-PLUS, EGRET,...) ofrecen grandes posibilidades en este sentido. Además de los gráficos vistos, es posible elaborar otros gráficos, incluso tridimensionales, permitiendo grandes cambios en su apariencia y facilidad de exportación a otros programas para presentar finalmente los resultados del estudio.

Figura 11. Dos diagramas de líneas superpuestos. Variación en el peso medio de una muestra de recién nacidos según el control ginecológico del embarazo y el hábito de fumar de la madre.

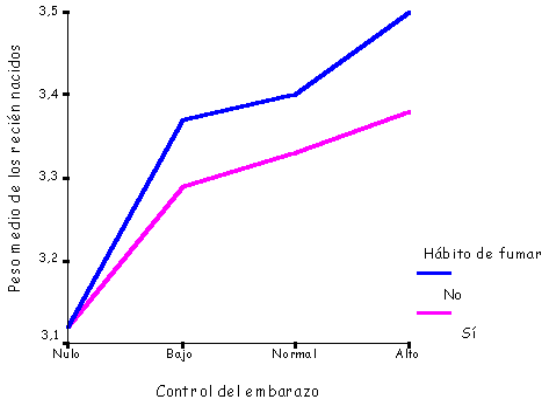


Figura 12. Diagrama de dispersión (regresión logística). Probabilidad de padecer cirrosis hepática, según un modelo de regresión logística ajustando por el % de protrombina y el presentar o no hepatomegalia.

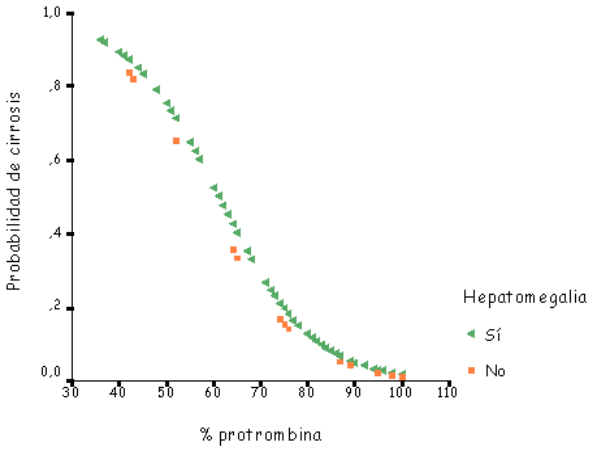
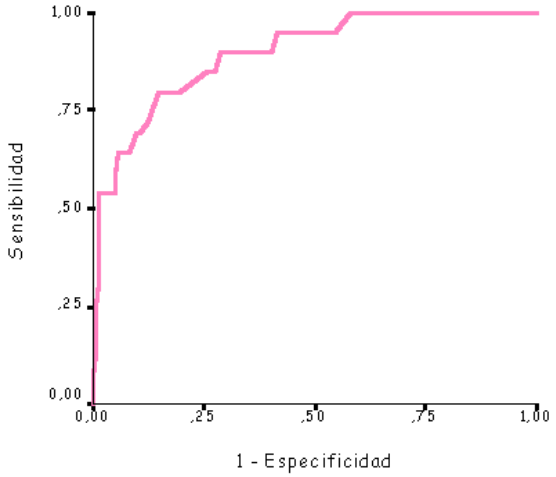


Figura 13. Curva ROC para el porcentaje de protrombina en la predicción de cirrosis.



2. PROBABILIDAD

La probabilidad es una medida de la certidumbre asociada a un suceso o evento futuro y suele expresarse como un número entre 0 y 1 (o entre 0 % y 100 %).

Una forma tradicional de estimar algunas probabilidades sería obtener la frecuencia de un acontecimiento determinado mediante la realización de experimentos aleatorios, de los que se conocen todos los resultados posibles, bajo condiciones suficientemente estables. Un suceso puede ser improbable (con probabilidad cercana a cero), probable (probabilidad intermedia) o seguro (con probabilidad uno).

La teoría de la probabilidad se usa extensamente en áreas como la estadística, la física, la matemática, las ciencias, la administración, contaduría, economía y la filosofía para sacar conclusiones sobre la probabilidad discreta de sucesos potenciales y la mecánica subyacente discreta de sistemas complejos, por lo tanto, es la rama de las matemáticas que estudia, mide o determina los experimentos o fenómenos aleatorios.

2.1. Espacio muestral

En la teoría de probabilidades, el espacio muestral o espacio de muestreo (denotado E, S, Ω o U) consiste en el conjunto de todos los posibles resultados de un experimento aleatorio, junto con una estructura sobre el mismo (ver más adelante).

Por ejemplo, si el experimento consiste en lanzar dos monedas, el espacio muestral es el conjunto {(cara, cara), (cara, cruz), (cruz, cara) y (cruz, cruz)}. Un evento o suceso es cualquier subconjunto del espacio muestral con estructura de σ-álgebra,1 llamándose a los sucesos que contengan un único elemento sucesos elementales. En el ejemplo, el suceso "sacar cara en el primer lanzamiento", o {(cara, cara), (cara, cruz)}, estaría formado por los sucesos elementales {(cara, cara)} y {(cara, cruz)}.

Para algunos tipos de experimento puede haber dos o más espacios de muestreo posibles. Por ejemplo, cuando se toma una carta de un mazo normal de 52 cartas, una posibilidad del espacio de muestreo podría ser el número (del as al rey), mientras que otra posibilidad sería el

palo (diamantes, tréboles, corazones y picas). Una descripción completa de los resultados, sin embargo, especificaría ambos valores, número y palo, y se podría construir un espacio de muestreo que describiese cada carta individual como el producto cartesiano de los dos espacios de muestreo descritos.

Los espacios de muestreo aparecen de forma natural en una aproximación elemental a la probabilidad, pero son también importantes en espacios de probabilidad. Un espacio de probabilidad (Ω, F, P) incorpora un espacio de muestreo de resultados, Ω, pero define un conjunto de sucesos de interés, la σ-álgebra F, por la cual se define la medida de probabilidad P.

Formalmente, un espacio muestral es una tripleta donde es el conjunto al que pertenecen los sucesos elementales, una colección de subconjuntos de que forma una σ-álgebra de subconjuntos (los subconjuntos , son los eventos aleatorios no elementales), y finalmente es una medida de conjuntos que permite asignar probabilidades a los sucesos o eventos del espacio muestral.

2.1.1. Tipos de espacio muestral

Podemos diferenciar entre dos tipos de espacios muestrales: discretos y continuos.

A. Discretos

Son aquellos espacios donde el número de sucesos elementales es finito o infinito numerable.

- ✓ **Espacio Probabilístico discreto**

Es aquel cuyo espacio muestral es discreto. Podemos diferenciar varios tipos de espacio probabilístico discreto:

- ✓ **Espacio Probabilístico Discreto Equiprobable**

Su espacio muestral es finito de tamaño n.

La probabilidad de cualquier suceso elemental E

- ✓ **Espacio Probabilístico Finito**

Su espacio muestral es discreto finito.

Hay al menos 2 sucesos elementales que cumplen.

- ✓ **Procesos Estocásticos Finitos Y Diagramas de Árbol**

Un proceso estocástico es una sucesión finita de experimentos aleatorios, cada uno de ellos con un nº finito de resultados posibles. Se representan con diagrama de árbol.

- ✓ **Espacio Probabilístico Infinito Contable**

Aquel cuyo espacio muestral es discreto infinito contable. Por ejemplo

La probabilidad de que salga cara en la primera tirada ----> ½

La probabilidad de que salga cara en la segunda tirada ---->½ * ½ = 1/4

La probabilidad de que salga cara en la tercera tirada ----> ½ * ½ * ½ = 1/8

B. Continuos

Son aquellos espacios donde el número de sucesos elementales es infinito incontable.

✓ **Espacio probabilístico continuo**

Espacio muestral infinito no numerable. -No es posible observar puntos concretos del espacio.

Tiene sentido hablar de intervalos observados. - No es posible asignar probabilidad a un punto concreto, se asigna a intervalos.

Por tanto la función P está definida sobre intervalos -----> P(Ki < Exp > Ke)

Habitualmente cuando trabajamos con magnitudes físicas.

✓ **Particiones**

Es posible definir particiones sobre el espacio muestral. Formalmente hablando, una partición sobre Ω se define como un conjunto numerable.

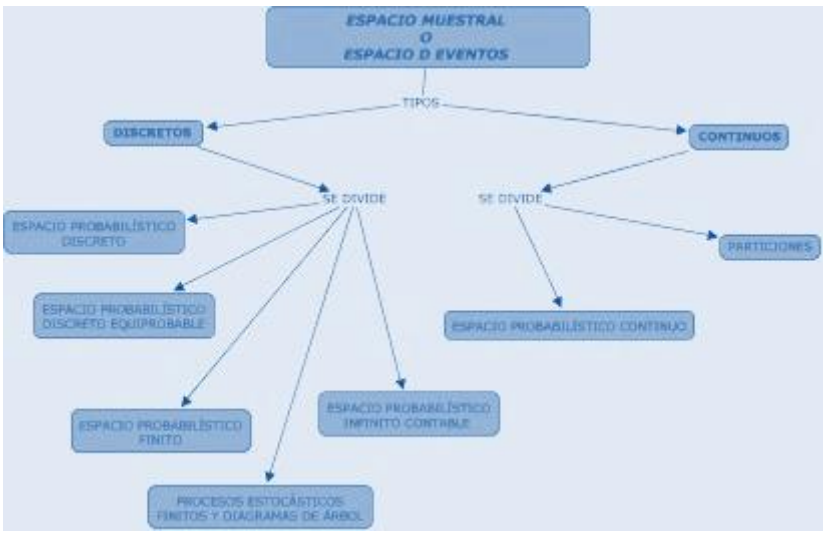
Ejemplos

Por ejemplo, en el caso del experimento aleatorio "lanzar un dado", el espacio muestral del experimento sería: Ω={1,2,3,4,5,6}. Por otro lado, si cambiamos ligeramente la experiencia pensando en el número resultante de la suma de 2 dados, entonces tenemos 2 espacios muestrales:

- Ω={(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1),(2,2),...(6,6)} = {1,2,3,4,5,6}x{1,2,3,4,5,6}
- Ω'={2,3,4,...,12}

La elección del espacio muestral es un factor determinante para realizar el cálculo de la probabilidad de un suceso.

En resumen, podemos mostrar lo siguiente



2.2. Eventos

En estadística, un evento o suceso es un subconjunto de un espacio muestral, es decir, un conjunto de posibles resultados que se pueden dar en un experimento aleatorio.

Formalmente, sea Ω un espacio muestral, entonces un evento es un subconjunto , donde (w1,w2,...) son una serie de posibles resultados.

Se dice que un evento A ocurre, si el resultado del experimento aleatorio es un elemento de A.

2.2.1. Tipos de eventos

A. Evento simple o suceso elemental

Un suceso o evento simple es un subconjunto del espacio muestral que contiene un único elemento.

Ejemplos de espacios muestrales y sucesos elementales:

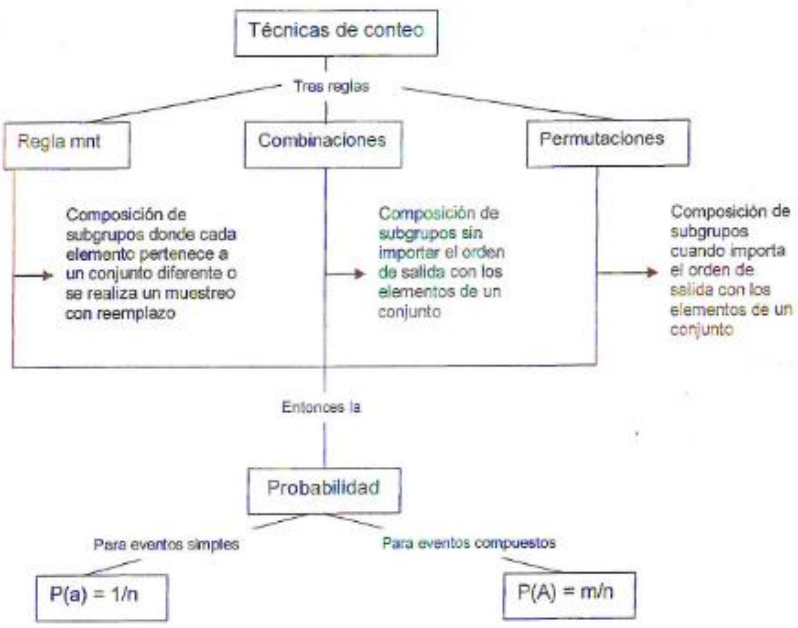
- ✓ Si se trata de contar objetos y el espacio muestral S = {0, 1, 2, 3, ...} (los números naturales), entonces los sucesos elementales son cada uno de los conjuntos {k}, donde k ∈ N.
- ✓ Si se lanza una moneda dos veces, S = {cc, cs, sc, ss}, donde (c representa "sale cara" y s, "sale cruz"), los sucesos elementales son {cc}, {cs}, {sc} y {ss}.
- ✓ Si X es una variable aleatoria normalmente distribuida, S = (-∞, +∞), los números reales, los sucesos elementalesson todos los conjuntos {x}, donde x ∈ .

Los sucesos elementales pueden tener probabilidades que son estrictamente mayores que cero, cero, no definidas o cualquier combinación de estas. Por ejemplo, la probabilidad de cualquier variable aleatoria discreta está determinada por las probabilidades asignadas a los sucesos elementales del experimento que determina la variable. Por otra parte, cualquier suceso elemental tiene probabilidad cero en cualquier variable aleatoria continua. Existen distribuciones mixtas que no son completamente continuas, ni completamente discretas, entre las que pueden darse ambas situaciones.

B. Otros sucesos

- Un evento compuesto es un subconjunto.
- Los eventos triviales son el conjunto universal Ω y el conjunto vacío. Al primero se le llama también evento seguro, y al segundo, evento imposible.
- Sean dos eventos A y B, si ambos son conjuntos disjuntos, entonces ellos son eventos excluyentes.
- Un evento con elementos infinitos pero numerables se llama σ-álgebra (sigma-álgebra), y un evento con elementos finitos se llama álgebra de sucesos de Boole.

2.3. Conteo de puntos muestrales



Son algunas reglas que se usan con el fin de contar el número de puntos muestrales de un experimento cuando estos no son fáciles de listar por su cantidad.

2.3.1. Regla MNT

Se usa cuando el experimento consiste en formar subconjuntos ordenados de elementos donde cada uno de estos elementos pertenece a un conjunto independiente o se hace un muestreo con reemplazo; por ejemplo el juego de lotería o chance; sacar una tripleta de un conjunto de 20 elementos donde un elemento puede ocupar las tres posiciones al tiempo, es decir el elemento que ha sido extraído vuelve a formar parte del conjunto inicial, esto es lo que se conoce como un muestreo con reemplazo.

2.3.2. Combinaciones

Se usa cuando el experimento consiste en formar subconjuntos de elementos de un mismo conjunto, es decir, un muestreo sin reemplazo, ademas no importa el orden de salida de los elementos; por ejemplo sacar de un grupo de 6 personas 3 para darles un premio, sacar de un grupo de 7 trabajadores 4 para realizar una misma tarea.

El número de combinaciones posibles esta dado por la siguiente expresión

$$C_r^n = \frac{n!}{(n-r)!\,xr!}$$

A. Tipos de Combinaciones

También hay dos tipos de combinaciones (recuerda que ahora el orden no importa):

- ✓ Se puede repetir: como monedas en tu bolsillo (5,5,5,10,10)
- ✓ Sin repetición: como números de lotería (2,14,15,27,30,33)

Combinaciones con repetición

En realidad, son las más difíciles de explicar, así que las dejamos para luego.

Combinaciones sin repetición

Así funciona la lotería. Los números se eligen de uno en uno, y si tienes los números de la suerte (da igual el orden) ¡entonces has ganado!

La manera más fácil de explicarlo es: Imaginemos que el orden sí importa (permutaciones), después lo cambiamos para que el orden no importe.

Volviendo a las bolas de billar, digamos que queremos saber qué 3 bolas se eligieron, no el orden.

Ya sabemos que 3 de 16 dan 3360 permutaciones.

Pero muchas de ellas son iguales para nosotros, porque no nos importa el orden.

Por ejemplo, digamos que se tomaron las bolas 1, 2 y 3. Las posibilidades son:

El orden importa	**El orden no importa**
1 2 3 1 3 2 2 1 3 2 3 1 3 1 2 3 2 1	1 2 3

Así que las permutaciones son 6 veces más posibilidades.

De hecho, hay una manera fácil de saber de cuántas maneras "1 2 3" se pueden ordenar, y ya la sabemos. La respuesta es:

$$\mathbf{3! = 3 \times 2 \times 1 = 6}$$

(Otro ejemplo: ¡4 cosas se pueden ordenar de 4! **= 4 × 3 × 2 × 1 = 24** maneras distintas, ¡prueba tú mismo!)

Así que sólo tenemos que ajustar nuestra fórmula de permutaciones para **reducir** por las maneras de ordenar los objetos elegidos (porque no nos interesa ordenarlos):

$$\frac{n!}{(n-r)!} \times \frac{1}{r!} = \frac{n!}{r!(n-r)!}$$

Esta fórmula es tan importante que normalmente se la escribe con grandes paréntesis, así:

$$\frac{n!}{r!(n-r)!} = \binom{n}{r}$$

donde **n** es el número de cosas que puedes elegir, y eliges **r** de ellas (No se puede repetir, el orden no importa) Y se la llama "coeficiente binomial".

Notación

Además de los "grandes paréntesis", la gente también usa estas notaciones:

$$C(n,r) = {}^{n}C_r = {}_{n}C_r = \binom{n}{r} = \frac{n!}{r!(n-r)!}$$

Ejemplo

Entonces, nuestro ejemplo de elegir 3 bolas de billar de 16 (ahora sin orden) es:

$$\frac{16!}{3!(16-3)!} = \frac{16!}{3!(13)!} = \frac{20{,}922{,}789{,}888{,}000}{6\times 6{,}227{,}020{,}800} = 560$$

O lo puedes hacer así:

$$\frac{16 \times 15 \times 14}{3 \times 2 \times 1} = \frac{3360}{6} = 560$$

Es interesante darse cuenta de que la fórmula es bonita y **simétrica**:

$$\frac{n!}{r!(n-r)!} = \binom{n}{r} = \binom{n}{n-r}$$

Con otras palabras, elegir 3 bolas de 16 da las mismas combinaciones que elegir 13 bolas de 16.

$$\frac{16!}{\mathbf{3}!\,(\mathbf{16}-\mathbf{3})!} = \frac{16!}{\mathbf{13}!\,(\mathbf{16}-\mathbf{13})!} = \frac{16!}{\mathbf{3}!\times\mathbf{13}!} = 560$$

Triángulo de Pascal

Puedes usar el triángulo de Pascal para calcular valores. Baja a la fila "n" (la de arriba es n=0), y ve a la derecha "r" posiciones, ese valor es la respuesta. Aquí tienes un trozo de la fila 16:

1 14 91 364 ...

1 15 105 455 1365 ...

1 16 120 560 1820 4368 ...

- ✓ **Combinaciones con repetición**

Digamos que tenemos cinco sabores de helado: **banana, chocolate, limón, fresa y vainilla**. Puedes tomar 3 paladas. ¿Cuántas variaciones hay?

Vamos a usar letras para los sabores: {b, c, l, f, v}. Algunos ejemplos son

- ✓ {c, c, c} (3 de chocolate)
- ✓ {b, l, v} (uno de banana, uno de limón y uno de vainilla)
- ✓ {b, v, v} (uno de banana, dos de vainilla)

(Y para dejarlo claro: hay n=5 cosas para elegir, y eliges r=3 de ellas. El orden no importa, ¡y sí puedes repetir!)

Bien, no puedo decirte directamente cómo se calcula, pero te voy a enseñar una técnica especial para que lo averigües tú mismo.

Imagina que el helado está en contenedores, podrías decir "sáltate el primero, después 3 paladas, después sáltate los 3 contenedores siguientes" ¡y acabarás con 3 paladas de chocolate!

Entonces es como si ordenaras a un robot que te trajera helado, pero no cambia nada, tendrás lo que quieres.

Ahora puedes escribirlo como →OOO→→→ (la flecha es saltar, el círculo es tomar)

Entonces los tres ejemplos de arriba se pueden escribir así:

- ✓ {c, c, c} (3 de chocolate): →OOO→→→
- ✓ {b, l, v} (uno de banana, uno de limón y uno de vainilla): O→→O→→O
- ✓ {b, v, v} (uno de banana, dos de vainilla): O→→→→OO

OK, entonces ya no nos tenemos que preocupar por diferentes sabores, ahora tenemos un problema **más simple** para resolver: "de cuántas maneras puedes ordenar flechas y círculos"

Fíjate en que siempre hay 3 círculos (3 paladas de helado) y 4 flechas (tenemos que movernos 4 veces para ir del contenedor 1º al 5º).

Así que (en general) hay ***r* + *(n-1)*** posiciones, y queremos que ***r*** de ellas tengan círculos.

Esto es como decir "tenemos ***r* + *(n-1)*** bolas de billar y queremos elegir ***r*** de ellas". Es decir, es como el problema de elegir bolas de billar, pero con números un poco distintos. Lo podrías escribir así:

$$\binom{n+r-1}{r} = \frac{(n+r-1)!}{r!(n-1)!}$$

donde n es el número de cosas que puedes elegir, y eliges r de ellas (Se puede repetir, el orden no importa)

Es interesante pensar que podríamos habernos fijado en flechas en vez de círculos, y entonces habríamos dicho "tenemos ***r* + *(n-1)*** posiciones y queremos que ***(n-1)*** tengan flechas", y la respuesta sería la misma...

$$\binom{n+r-1}{r} = \binom{n+r-1}{n-1} = \frac{(n+r-1)!}{r!(n-1)!}$$

¿Qué pasa con nuestro ejemplo, cuál es la respuesta?

$$\frac{(\mathbf{5}+\mathbf{3}-\mathbf{1})!}{\mathbf{3}!\,(\mathbf{5}-\mathbf{1})!} = \frac{\mathbf{7}!}{\mathbf{3}!\times\mathbf{4}!} = \frac{\mathbf{5040}}{\mathbf{6}\times\mathbf{24}} = 35$$

En conclusión es un montón de cosas que absorber, quizás tendrías que leerlo otra vez para entenderlo todo bien!

Pero saber *cómo* funcionan estas fórmulas es sólo la mitad del trabajo. Averiguar cómo se interpreta una situación real puede ser bastante complicado.

Por lo menos ahora sabes cómo se calculan las 4 variantes de "el orden sí/no importa" y "sí/no se puede repetir".

2.3.3. Permutaciones

Se usa cuando el experimento consiste en formar subconjuntos de elementos de un mismo conjunto, es decir, un muestreo sin reemplazo, pero en este caso si importa el orden de salida de los elementos; por ejemplo sacar de un grupo de 6 personas 3 para darles un premio, pero dependiendo de su orden de salida el premio será mejor. Sacar de un grupo de 7 trabajadores 4 para realizar una tarea distinta cada uno.

El número de permutaciones posibles está dado por la siguiente expresión:

$$P_r^n = \frac{n!}{(n-r)!}$$

Nótese que

$$P_r^n = C_r^n x r!$$

Nota: Recordar que P(A) = m/n

$$m = C_{r-x}^{n-x} x r!$$

Donde x es el número de elementos escogidos del subgrupo r. (Si importa el orden).

$$m = C_{r-x}^{n-x}$$

Donde x es el número de elementos escogidos del subgrupo r. (Si no importa el orden).

n: número de combinaciones o permutaciones según el caso.

A. ¿Qué diferencia hay Combinaciones y permutaciones?

Normalmente usamos la palabra "combinación" descuidadamente, sin pensar en si el orden de las cosas es importante. En otras palabras:

"Mi ensalada de frutas es una combinación de manzanas, uvas y bananas": no importa en qué orden pusimos las frutas, podría ser "bananas, uvas y manzanas" o "uvas, manzanas y bananas", es la misma ensalada.

"La combinación de la cerradura es 472": ahora sí importa el orden. "724" no funcionaría, ni "247". Tiene que ser exactamente 4-7-2.

Así que en matemáticas usamos un lenguaje más preciso:

- ✓ Si el orden no importa, es una **combinación**.
- ✓ Si el orden **sí** importa es una **permutación**.

Con otras palabras: Una permutación es una combinación **ordenada**.

B. Tipos de Permutaciones

Hay dos tipos de permutaciones:

- ✓ **Se permite repetir**: como la cerradura de arriba, podría ser "333".
- ✓ **Sin repetición**: por ejemplo los tres primeros en una carrera. No puedes quedar primero *y* segundo a la vez.

- ✓ **Permutaciones con repetición**

Son las más fáciles de calcular. Si tienes *n* cosas para elegir y eliges *r* de ellas, las permutaciones posibles son:

$$n \times n \times \ldots \text{ (r veces)} = n^r$$

(Porque hay n posibilidades para la primera elección, DESPUÉS hay n posibilidades para la segunda elección, y así.)

Por ejemplo, en la cerradura de arriba, hay 10 números para elegir (0, 1, ...,9) y eliges 3 de ellos:

$$10 \times 10 \times \ldots \text{ (3 veces)} = 10^3 = 1000 \text{ permutaciones}$$

Así que la fórmula es simplemente: $\mathbf{n^r}$ donde ***n*** es el número de cosas que puedes elegir, y eliges ***r*** de ellas (Se puede repetir, el orden importa)

- ✓ **Permutaciones sin repetición**

En este caso, se reduce el número de opciones en cada paso.

Por ejemplo, ¿cómo podrías ordenar 16 bolas de billar? Después de elegir por ejemplo la "14" no puedes elegirla otra vez.

Así que tu primera elección tiene 16 posibilidades, y tu siguiente elección tiene 15 posibilidades, después 14, 13, etc. Y el total de permutaciones sería:

$$16 \times 15 \times 14 \times 13 \ldots = 20{,}922{,}789{,}888{,}000$$

Pero a lo mejor no quieres elegirlas todas, sólo 3 de ellas, así que sería solamente:

$$16 \times 15 \times 14 = 3360$$

Es decir, hay 3,360 maneras diferentes de elegir 3 bolas de billar de entre 16. ¿Pero cómo lo escribimos matemáticamente? Respuesta: usamos la "función factorial"

La función factorial (símbolo: !) significa que se multiplican números descendentes. Ejemplos:

- ✓ $4! = 4 \times 3 \times 2 \times 1 = 24$
- ✓ $7! = 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 5040$
- ✓ $1! = 1$

Nota: en general se está de acuerdo en que 0! = 1. Puede que parezca curioso que no multiplicar ningún número dé 1, pero ayuda a simplificar muchas ecuaciones.

Así que si quieres elegir todas las bolas de billar las permutaciones serían:

$$16! = 20{,}922{,}789{,}888{,}000$$

Pero si sólo quieres elegir 3, tienes que dejar de multiplicar después de 14. ¿Cómo lo escribimos? Hay un buen truco... dividimos entre 13!...

$$\frac{\mathbf{16 \times 15 \times 14 \times 13 \times 12 \ldots}}{\mathbf{13 \times 12 \ldots}} = \mathbf{16 \times 15 \times 14 = 3360}$$

¿Lo ves? 16! / 13! = 16 × 15 × 14

La fórmula se escribe: $\frac{n!}{(n-r)!}$ donde n es el número de cosas que puedes elegir, y eliges r de ellas (No se puede repetir, el orden importa)

Notación

En lugar de escribir toda la fórmula, la gente usa otras notaciones como:

$$P(n,r) = {}^{n}P_r = {}_{n}P_r = \frac{n!}{(n-r)!}$$

2.4. Probabilidad de un evento

Algunas situaciones de probabilidad implican más de un evento. Cuando los eventos no se afectan entre sí, se les conoce como eventos independientes. Los eventos independientes pueden incluir la repetición de una acción como lanzar un dado más de una vez, o usar dos elementos aleatorios diferentes, como lanzar una moneda y girar una ruleta. Muchas otras situaciones también pueden incluir eventos independientes. Para calcular correctamente las probabilidades, necesitamos saber si un evento influye en el resultado de otros eventos.

2.4.1. Cálculo o medición de la Probabilidad

La probabilidad mide la mayor o menor posibilidad de que se dé un determinado resultado (suceso o evento) cuando se realiza un experimento aleatorio.

Para calcular la probabilidad de un evento se toma en cuenta todos los casos posibles de ocurrencia del mismo; es decir, de cuántas formas puede ocurrir determinada situación.

Los casos favorables de ocurrencia de un evento serán los que cumplan con la condición que estamos buscando.

La probabilidad toma valores entre 0 y 1 (o expresados en tanto por ciento, entre 0% y 100%):

El **valor cero** corresponde al suceso imposible; ejemplo: lanzamos un dado al aire y la probabilidad de que salga el número 7 es cero.

El **valor uno** corresponde al suceso seguro, ejemplo: lanzamos un dado al aire y la probabilidad de que salga cualquier número del 1 al 6 es igual a uno (100%).

El resto de sucesos tendrá probabilidades entre cero y uno: que será tanto mayor cuanto más probable sea que dicho suceso tenga lugar.

Métodos de medición de Probabilidad

Uno de los métodos más utilizados es aplicando la Regla de Laplace: define la probabilidad de un suceso como el cociente entre casos favorables y casos posibles.

$$P_{(suceso)} = \frac{\text{casos favorables (f)}}{\text{casos posibles (n)}}$$

Ejemplos:

a) Probabilidad de que al lanzar un dado salga el número 2: el caso favorable (f) es tan sólo uno (que salga el dos), mientras que los casos posibles (n) son seis (puede salir cualquier número del uno al seis).

Por lo tanto:

$$P_{(suceso)} = \frac{f}{n} = \frac{1}{6} = 0{,}166$$

(o lo que es lo mismo, 16,6%)

b) Probabilidad de que al lanzar un dado salga un número par: en este caso los casos favorables (f) son tres (que salga el dos, el cuatro o el seis), mientras que los casos posibles (n) siguen siendo seis.

Por lo tanto:

$$P_{(suceso)} = \frac{f}{n} = \frac{3}{6} = \frac{1}{2} = 0{,}50$$ (o lo que es lo mismo, 50%)

c) Probabilidad de que al lanzar un dado salga un número menor que 5: en este caso tenemos cuatro casos favorables (f) (que salga el uno, el dos, el tres o el cuatro), frente a los seis casos posibles.

Por lo tanto:

$$P_{(suceso)} = \frac{f}{n} = \frac{4}{6} = \frac{2}{3} = 0{,}666$$ (o lo que es lo mismo, 66,6%)

d) Probabilidad de ganarse el premio mayor de una lotería en la que juegan 100.000 númerosnos: tan sólo un caso favorable (f), el número que jugamos, frente a los 100.000 casos posibles (n).

Por lo tanto:

$$P_{(suceso)} = \frac{1}{100.000} = 0{,}00001$$ (o lo que es lo mismo, 0,001%)

d) Probabilidad al lanzar una moneda, con un águila en una cara y un sol en la otra. Hay dos casos posibles (n) de ocurrencia (o cae águila o cae sol) y sólo un caso favorable (f) de que pueda caer águila (pues sólo hay un águila en la moneda).

Por lo tanto:

$$P_{(águila)} = \frac{f}{n} = \frac{1}{2} = 0{,}5$$ (o, lo que es lo mismo, 50 %)

Existe una probabilidad del 50% de obtener un águila al tirar una moneda.

e) Probabilidad de elegir tal o cual fruta. Si en una canasta hay 20 peras y 10 manzanas. ¿Qué fruta es más probable que saque al azar de la canasta?

Para este ejemplo tenemos que 30 es el total de frutas en la canasta; es decir los casos posibles (n). Para calcular la probabilidad de sacar una manzana los casos favorables (f) son 10 puesto que existen sólo 10 manzanas.

Por lo tanto:

$$P_{(manzana)} = \frac{f}{n} = \frac{10}{30} = \frac{1}{3} = 0{,}333$$ (o, lo que es lo mismo, 33,3 %)

$$P_{(pera)} = \frac{f}{n} = \frac{20}{30} = \frac{2}{3} = 0{,}667$$ (o, lo que es lo mismo, 66,7 %)

Fíjate bien que 33,3% + 66,7% es igual al 100% porque siempre que saquemos algo de la canasta es seguro que será una fruta.

2.4.2. Condiciones importantes

Para poder aplicar la Regla de Laplace el experimento aleatorio tiene que cumplir dos requisitos:

a) El número de resultados posibles (sucesos o eventos) tiene que ser finito. Si hubiera infinitos resultados, al aplicar la regla "casos favorables dividido por casos posibles" el cociente siempre sería cero.

b) Todos los sucesos o eventos tienen que tener la misma probabilidad. Si al lanzar un dado, algunas caras tuvieran mayor probabilidad de salir que otras, no podríamos aplicar esta regla.

A la regla de Laplace también se le denomina "probabilidad a priori", ya que para aplicarla hay que conocer antes de realizar el experimento cuales son los posibles resultados y saber que todos tienen las mismas probabilidades.

Cuando se realiza un experimento aleatorio un número muy elevado de veces, las probabilidades de los diversos posibles sucesos empiezan a converger hacia valores determinados, que son sus respectivas probabilidades.

Ejemplo:

si lanzo una vez una moneda al aire y sale "cara", quiere decir que el suceso "cara" ha aparecido el 100% de las veces y el suceso "cruz" el 0%.

Si lanzo diez veces la moneda al aire, es posible que el suceso "cara" salga 7 veces y el suceso "cruz" los 3 restantes. En este caso, la probabilidad del suceso "cara" ya no sería del 100%, sino que se habría reducido al 70%.

Si repito este experimento un número elevado de veces, lo normal es que las probabilidades de los sucesos "cara" y "cruz" se vayan aproximando al 50% cada una. Este 50% será la probabilidad de estos sucesos según el modelo frecuentista.

2.5. Reglas aditivas

A menudo resulta más sencillo calcular la probabilidad de algún evento a partir del conocimiento de las probabilidades de otros eventos. Esto puede ser cierto si el evento en cuestión se puede representar como la unión de otros dos eventos o como el complemento de algún evento. Para simplificar cálculos se crean reglas como la regla aditiva, se aplica a uniones de eventos.

Si A y B son dos eventos, entonces

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Se clasifican en:

2.5.1. Eventos no Mutuamente Excluyentes

Dos o más eventos son no excluyentes o conjuntos, cuando es posible que ocurran ambos. Esto no indica que necesariamente deban ocurrir estos eventos en forma simultánea.

Ejemplo: Si consideramos en un juego de domino sacar al menos un blanco y un seis, estos eventos son no excluyentes porque puede ocurrir que salga el seis blanco.

2.5.2. Eventos Mutuamente Excluyentes

Dos o más eventos son mutuamente excluyentes o disjuntos, si no pueden ocurrir simultáneamente. Es decir, la ocurrencia de un evento impide automáticamente la ocurrencia del otro evento (o eventos).

Ejemplo: Al lanzar una moneda solo puede ocurrir que salga cara o sello, pero no los dos a la vez, esto quiere decir que estos eventos son excluyentes.

2.5.3. Eventos Complementarios

Los eventos complementarios son dos resultados de un evento, siendo éstos los *dos únicos resultados posibles*.

Ejemplo: Es como lanzar una moneda y que salga cara o cruz. Claro, no hay más opciones, así que estos eventos son complementarios.

Si A y B son dos eventos, entonces $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Ejemplos:

La probabilidad de que Paula apruebe Estadística es 2/3 y la probabilidad de que apruebe Ecuaciones Diferenciales es 4/9. Si la probabilidad de aprobar ambos cursos es 1/4. ¿Cuál es la probabilidad de que Paula apruebe al menos uno de estos cursos? Si E es el evento aprobar Estadística y D el evento aprobar Ecuaciones Diferenciales, entonces

$$P(E \cup D) = P(E) + P(D) - P(E \cap D) = \frac{2}{3} + \frac{4}{9} - \frac{1}{4} = \frac{31}{36}$$

¿Cuál es la probabilidad de obtener un total de siete u once cuando se lanza un par de dados? Sea A el evento de que ocurre siete y B el evento de que salga once. Como todos los puntos muestrales son igualmente probables, P(A)=1/6 y P(B)=1/18. Los eventos A y B son mutuamente excluyentes, pues un total de 7 y 11 no pueden ocurrir en el mismo lanzamiento.

$$P(A \cup B) = P(A) + P(B) = \frac{1}{6} + \frac{1}{18} = \frac{2}{9}$$

Si las probabilidades de que una persona que compra un automóvil nuevo elija el color verde, blanco, rojo o azul son, respectivamente, 0.09, 0.15, 0.21 y 0.23. ¿Cuál es la probabilidad de que un comprador dado adquiera un automóvil nuevo que tenga uno de esos colores? Sean V, B, R y A los eventos de que un comprador seleccione, respectivamente, un automóvil verde, blanco, rojo o azul. Como estos eventos son mutuamente excluyentes

$$\begin{aligned} P(V \cup B \cup R \cup A) &= P(V) + P(B) + P(R) + P(A) \\ &= 0.09 + 0.15 + 0.21 + 0.23 \\ &= 0.68 \end{aligned}$$

A veces es más difícil calcular la probabilidad de que ocurra un evento que calcular la probabilidad de que el evento no ocurra. Si este es el caso para algún evento A, simplemente calculamos primero P(A') y después, usamos lo siguiente, para calcular P(A) por sustracción.

• Si A y A' son eventos complementarios, entonces

$$P(A) + P(A') = 1$$

Si las probabilidades de que un mecánico automotriz de servicio a tres, cuatro, cinco, seis, siete, ocho o más autos en un día de trabajo son 0.12, 0.19, 0.28, 0.24, 0.10 y 0.07, respectivamente. ¿Cuál es la probabilidad de que de servicio al menos a cinco autos el día siguiente de trabajo? Sea S el evento de que al menos cinco autos reciban servicio. Luego P(S)=1-P(S') donde S' es el evento de que menos de cinco autos reciban servicio. Como P(S')=0.12+0.19=0.31, entonces P(S)=1-0.31=0.69.

2.6. Probabilidad condicional

Probabilidad condicional es la probabilidad de que ocurra un evento A, sabiendo que también sucede otro evento B. La probabilidad condicional se escribe P(A|B) o P(A/B), y se lee «la probabilidad de A dado B».

No tiene por qué haber una relación causal o temporal entre A y B. A puede preceder en el tiempo a B, sucederlo o pueden ocurrir simultáneamente. A puede causar B, viceversa o pueden no tener relación causal. Las relaciones causales o temporales son nociones que no pertenecen al ámbito de la probabilidad. Pueden desempeñar un papel o no, dependiendo de la interpretación que se le dé a los eventos.

Un ejemplo clásico es el lanzamiento de una moneda para luego lanzar un dado. ¿Cuál es la probabilidad que en el dado salga un 6 dado que ya haya salido una cara en la moneda? Esta probabilidad se denota de esta manera: P(6|C).

El condicionamiento de probabilidades puede lograrse aplicando el teorema de Bayes.

Como la probabilidad está ligada a nuestra ignorancia sobre los resultados de la experiencia, el hecho de que ocurra un suceso, puede cambiar la probabilidad de los demás. El proceso de realizar la historia clínica, explorar y realizar pruebas complementarias ilustra este principio.

La probabilidad de que ocurra el suceso A si ha ocurrido el suceso B se denomina probabilidad condicionada y se define

$$p(A|B) = \frac{p(A \cap B)}{p(B)} \quad \text{si } p(B) \neq 0$$

Esta definición es consistente, es decir cumple los axiomas de probabilidad.

Cuando ocurre un suceso cambia el espacio muestral, por eso cambia la probabilidad. A veces es más fácil calcular la probabilidad condicionada teniendo en cuenta este cambio de

Ejemplo : Supongamos que nuestro EM es una muestra de adultos de que cumplen con los requisitos para obtener un grado en la Facultad de Ingeniería. Debemos clasificarlos de acuerdo con su género y situación laboral.

	Empleado	Desempleado	Total
Hombre	460	40	500
Mujer	140	260	400
Total	600	300	900

Uno de estos individuos se seleccionará al azar para que realice un viaje a través del país para promover las ventajas de establecer industrias nuevas en Huancayo. Nos interesan los siguientes eventos:

H: se elige un hombre, E: el seleccionado tiene empleo.

Al utilizar el EM reducido E, tenemos

$$P(H|E) = \frac{460}{600} = \frac{23}{30}$$

Para verificar este resultado, notar que

$$P(E) = \frac{600}{900} = \frac{2}{3} \quad \text{y} \quad P(E \cap H) = \frac{460}{900} = \frac{23}{45}$$

$$P(H|E) = \frac{23/45}{2/3} = \frac{23}{30}$$

2.7. Reglas multiplicativas

Si en un experimento pueden ocurrir los eventos A y B, entonces

$P(A \cap B) = P(A)P(B|A)$, dado que $P(A)>0$.

Así la probabilidad de que ocurran A y B es igual a la probabilidad de que ocurra A multiplicada por la probabilidad condicional de que ocurra B, dado que ocurre A.

Como los eventos $A \cap B$ y $B \cap a$ son equivalentes, del teorema anterior se sigue que también podemos escribir

$P(A \cap B) = P(B \cap A) = P(B)P(A|B)$.

En otras palabras, no importa qué evento se considere como A y cuál como B.

Ejemplo: Suponga que tenemos una caja de fusibles que contiene 20 unidades, de las cuales 5 están defectuosas. Si se seleccionan 2 fusibles al azar y se retiran de la caja, uno después del otro, sin reemplazar el primero, ¿cuál es la probabilidad de que ambos fusibles estén defectuosos?

Sean A el evento de que el primer fusible esté defectuoso y B ele vento de que el segundo esté defectuoso; entonces, interpretamos $A \cap B$ como el evento de que ocurra A, y entonces B ocurre después de que haya ocurrido A. La probabilidad de separar primero un fusible defectuoso es 1/4; entonces, la probabilidad de separar un segundo fusible defectuoso de los restantes 4 es 4/19. Por lo tanto

$P(A \cap B) = (1/4)(4/19) = 1/19$.

2.7.1. Eventos independientes

Dos eventos A y B son independientes si y sólo si

$$P(A \cap B) = P(A)P(B).$$

Por lo tanto, para obtener la probabilidad de que ocurran dos eventos independientes, simplemente calculamos el producto de sus probabilidades individuales.

2.8. Regla de Bayes

Si los sucesos A_i s**on una partición** y B un suceso tal que $p(B) \neq 0$

$$p(A_i|B) = \frac{p(B|A_i)p(A_i)}{\sum_{j=1}^{n} p(B|A_j)p(A_j)} \quad \text{para } i = 1,\ldots n$$

Aplicaciones

Diagnóstico médico (en general clasificaciones no biunívocas): El diagnóstico consiste en establecer la enfermedad de un paciente, a partir de una serie de síntomas. Pero los síntomas y las enfermedades no están ligados de un modo biunívoco.

Llamemos Ei al conjunto de enfermedades E1: tuberculosis pulmonar; E2 :cáncer de pulmón; E3: bronquitis obstructiva; etc. y Si a los síntomas y síndromes asociados con las mismas. S1: tos; S2: estado febril; S3: hemotisis; etc.

La información accesible en los libros de patología, o en un archivo de historias clínicas es del tipo.

Para E1: algunos (digamos el 20%) tienen hemotisis; muchos (80%) tienen tos; etc. y lo mismo para las demás enfermedades.

En términos de probabilidad condicionada, esta información es p(S3|E1) = 0,2; p(S1|E1) = 0,8 etc.

para diagnosticar la tuberculosis se ha de evaluar, para los síntomas que presenta el paciente p(E1|Si) para lo que se puede usar el teorema de Bayes si las enfermedades forman una partición (son mutuamente excluyentes y se consideran todas las enfermedades compatibles con el síntoma) y se conocen sus prevalencias.

Nótese que un mismo conjunto de síntomas podría dar lugar a un diagnóstico diferente en poblaciones en las que las prevalencias fueran diferentes.

Pruebas diagnósticas: Supóngase una prueba diagnóstica, por ejemplo nivel de glucosa en sangre, en ayunas, para diagnosticar la diabetes. Se considera que la prueba es positiva si se encuentra un nivel por encima de un cierto valor, digamos 120 mg/l.

Para evaluar la prueba, (habrá que hacerlo para distintos valores de corte) se somete a la misma a una serie de individuos diabéticos diagnosticados por otro procedimiento (el patrón de oro o "gold standar") y a una serie de individuos no diabéticos. Los resultados se pueden representar en una tabla de doble entrada.

		Patrón de oro		
		NE	E	
Prueba	-	a	b	r
	+	c	d	s
		t	u	

Si la prueba fuera perfecta b=c=0, desgraciadamente nunca ocurre. Se denomina coeficiente falso-positivo (CFP) al cociente c/t, y es una estimación de la probabilidad condicionada p(+|NE), se denomina coeficiente falso-negativo (CFN) al cociente b/u, y es una estimación de la probabilidad condicionada p(-|E). Estos dos coeficientes cuantifican los dos errores que la prueba puede cometer y caracterizan a la misma. Simétricamente, los coeficientes que cuantifican los aciertos son la sensibilidad, p(+|E), y la especificidad p(-|NE).

Cuando la prueba se usa con fines diagnósticos (o de "screening") interesa calcular p(E|+) y/o p(NE|-).

Como E y NE son una partición, usando el Teorema de Bayes

$$p(E|+) = \frac{p(+|E)p(E)}{p(+|E)p(E) + p(+|NE)p(NE)} \quad \text{y} \quad p(NE|-) = \frac{p(-|NE)p(NE)}{p(-|E)p(E) + p(-|NE)p(NE)}$$

Nótese que ambas dependen de la prevalencia de la enfermedad: una prueba diagnóstica que funciona muy bien en la clínica Mayo, puede ser inútil en el Hospital Ramón y Cajal.

Ejemplo 9:

una prueba diagnóstica para la diabetes tiene un CFP de 4% y un CFN del 5%. Si la prevalencia de la diabetes en la población donde se usa es del 7% ¿cuál es la probabilidad de que sea diabético un individuo en el que la prueba dé positiva? y ¿de que no lo sea uno en el que dé negativo?

p(+|NE) = 0,04 ⇒ p(-|NE) = 0,96

p(-|E) = 0,05 ⇒ p(+|E) = 0,95

p(E) = 0,07 ⇒ p(NE) = 0,93

$$p(E|+) = \frac{0{,}95 \times 0{,}07}{0{,}95 \times 0{,}07 + 0{,}04 \times 0{,}93} = 0{,}641$$

y

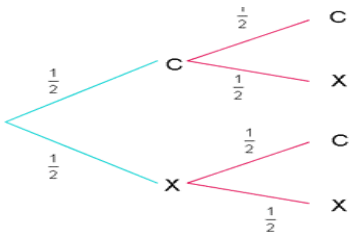
$$p(NE|-) = \frac{0{,}96 \times 0{,}93}{0{,}05 \times 0{,}07 + 0{,}96 \times 0{,}93} = 0{,}996$$

Pruebas en serie: Cuando se aplican pruebas en serie, para cada prueba p(E) y p(NE), serán la p(E|+) y p(NE|+) de la prueba anterior (si dio positiva) o p(E|-) y p(NE|-) si dio negativa.

2.9. Ejercicios de Repaso

Hallar la probabilidad de que al lanzar al aire dos monedas, salgan:

1. Dos caras.



Son sucesos independientes

$$p(2c) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

2. Dos cruces.

Son sucesos independientes

$$p(2\times) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

3. Una cara y una cruz.

La probabilidad de sacar una cara y una cruz sería la probabilidad de cx o de xc:

$$P(c \cap x) \cup P(x \cap c)$$

$$p(1c \cap 1x) = \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2}$$

4. Hallar la probabilidad de que al levantar unas fichas de dominó se obtenga un número de puntos mayor que 9 o que sea múltiplo de 4.

$$A(>9) = \{(4,6),(5,5),(5,6),(6,6)\}$$
$$B(4) = \{(b,4),(1,3),(2,2),(2,6),(3,5),(4,4),(6,6)\}$$

Vemos que son sucesos compatibles porque (6,6) es común, por tanto $A \cap B \neq \emptyset$

$$p(A \cup B) = \frac{4}{28} + \frac{7}{28} - \frac{1}{28} = \frac{10}{28} = \frac{5}{14}$$

5. Se lanzan dos dados al aire y se anota la suma de los puntos obtenidos. Se pide:

La probabilidad de que salga el 7

1	2	3	4	5	6
6	5	4	3	2	1

$$p(7) = \frac{6}{36} = \frac{1}{6}$$

La probabilidad de que el número obtenido sea par

$$p(par) = \frac{18}{36} = \frac{1}{2}$$

La probabilidad de que el número obtenido sea múltiplo de tres

1	1	2	2	3	3	4	4	5	5	6	6
2	5	1	4	3	6	2	5	1	4	3	6

$$p(3) = \frac{12}{36} = \frac{1}{3}$$

6. Se extraen cinco cartas de una baraja de 52. Hallar la probabilidad de extraer:

4 ases

$$p(4 \text{ ases}) = \frac{C_4^4 \cdot C_{48}^1}{C_{52}^5} = \frac{1}{54145}$$

4 ases y un rey

$$p(4 \text{ ases y un rey}) = \frac{C_4^4 \cdot C_4^1}{C_{52}^5} = \frac{1}{649740}$$

3 cincos y 2 sotas

$$p(3 \text{ cincos y dos sotas}) = \frac{C_4^3 \cdot C_4^2}{C_{52}^5} = \frac{1}{108290}$$

Un 9, 10, sota, caballo y rey en cualquier orden

$$p(escalera) = \frac{C_4^1 \cdot C_4^1 \cdot C_4^1 \cdot C_4^1 \cdot C_4^1}{C_{52}^5} = \frac{64}{162435}$$

3 de un palo cualquiera y 2 de otro

$$p(3\ y\ 2) = \frac{4 \cdot C_{13}^{3} \cdot 3 \cdot C_{13}^{2}}{C_{52}^{5}} = \frac{429}{4165}$$

Al menos un as

$$p(\text{ningún as}) = \frac{C_{48}^{5}}{C_{52}^{5}} = \frac{35673}{54145}$$

3. VARIABLES ALEATORIAS Y DISTRIBUCIONES DE PROBABILIDAD

3.1. Concepto de variable aleatoria

Una función que asocia un número real, perfectamente definido, a cada punto muestral.

A veces las variables aleatorias (v.a.) están ya implícitas en los puntos muestrales.

Ejemplo 1: Experiencia consistente en medir la presión sistólica de 100 individuos. Un punto muestral (resultado de un experimento) es ya un número (presión sistólica). La v.a. está implícita.

Ejemplo : En el ejemplo de la mujer portadora de hemofilia.

mujer portadora de hemofilia = {*sss, ssn, sns, snn, nss, nsn, nns, nnn*}

Se podría definir una variable que asignara a cada punto muestral el número de orden en el espacio muestral.

X: *sss* $\rightarrow$ 1; *ssn* $\rightarrow$ 2; *sns* $\rightarrow$ 3;...

Pero otra posible v.a.: a cada punto muestral el número de *s*. X: *sss* $\rightarrow$ 3; *ssn* $\rightarrow$ 2; ...

Los conjuntos pueden ser:

- ✓ *discretos:* número finito o infinito numerable de elementos.
- ✓ *continuos:* número infinito no numerable de elementos.

Las v.a. definidas sobre espacios muestrales discretos se llaman v.a. discretas y las definidas sobre espacios muestrales continuos se llaman continuas.

Una v.a. puede ser continua, aunque nosotros sólo podamos acceder a un subconjunto finito de valores. P.e. la presión arterial es una v.a. continua pero sólo podemos acceder a un conjunto finito de valores por la limitación de los aparatos de medida.

En general, las *medidas* dan lugar a v.a. continuas y los *conteos* a v.a. discretas.

Inducción de la probabilidad a variables aleatorias

Las v.a permiten definir la probabilidad como una función numérica (de variable real) en lugar de como una función de conjunto como se había definido antes

Ejemplo 3: Tiramos una moneda 3 veces. Representamos cara por c y cruz por z.

Tiramos una moneda 3 veces = {ccc, ccz, czc, zcc, czz, zcz, zzc, zzz}

La probabilidad de cada suceso elemental es 1/8. Por ejemplo p(ccc)=1/8, ya que la probabilidad de sacar cara en una tirada es 1/2 según la definición clásica y las tiradas son independientes.

Definimos la v.a. X: número de caras, que puede tomar los valores {0, 1, 2, 3}. Se buscan todos los puntos muestrales que dan lugar a cada valor de la variable y a ese valor se le asigna la probabilidad del suceso correspondiente.

x	Sucesos	Px
0	{zzz}	1/8
1	{czz, zcz, zzc}	3/8
2	{ccz, czc, zcc}	3/8
3	{ccc}	1/8

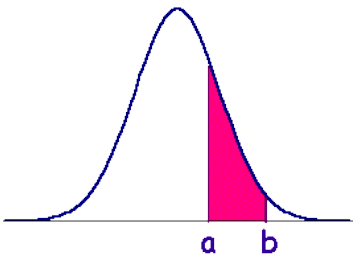
A esta función se le denomina función densidad de probabilidad (fdp), que desgraciadamente "funciona" de distinta manera en las variables discreta que en las continuas. En el caso de las variables discretas, como en el ejemplo, es una función que para cada valor de la variable da su probabilidad.

Ejemplo : Supongamos la variable tipo histológico de un tumor, con los valores 1, 2, 3, 4. Si la fdp fuera

x	f(x)
1	0,22
2	0,27
3	0,30
4	0,21

significaría que la probabilidad del tipo 2 es 0,27, etc.

Para variables continuas la probabilidad de que una variable tome cualquier valor concreto es 0, por lo tanto la fdp sólo permite calcular la probabilidad para un intervalo del tipo (a<X<b), mediante el área bajo la curva de la fdp.



Parámetros característicos de una *fdp*

Valor esperado o esperanza matemática o media

$$\mu_x = E[x] = \begin{cases} \sum_x xf(x) & \text{caso dis.} \\ \int_{-\infty}^{\infty} xf(x)\,dx & \text{caso cont.} \end{cases}$$

si X es una v.a. cualquier función de ella, h(x), es también una v.a., en consecuencia, también se define este parámetro para una función de v.a.

$$\mu_h = E[h(x)] = \begin{cases} \sum_x h(x)\, f(x) & \text{c. d.} \\ \int_{-\infty}^{\infty} h(x)\, f(x)\, dx & \text{c. c.} \end{cases}$$

Ejemplo: Se tira un dado. Se define como v.a. el número que sale ¿Cuál es su media?

La variable X puede tomar los valores 1, 2, ..., 6 y para todos ellos f(x) = 1/6. En consecuencia, la media es

$$\mu_x = \sum_{x=1}^{6} x f(x) = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + \ldots + 6 \times \frac{1}{6} = 3,5$$

Obsérvese que es un número que la v.a. no puede alcanzar. ¿Qué significa? No mucho.

Se define ahora una función sobre X: el premio: si sale 1 ó 2 se gana 100 ptas, si sale 3 se gana 500 y si sale 4, 5 ó 6 no se gana nada

X	h(x)
1	100
2	100
3	500
4	0
5	0
6	0

¿Cuál es el valor medio de esta función?

$$\mu_x = \sum_{x=1}^{6} h(x) f(x) = 100 \times \frac{1}{6} + 100 \times \frac{1}{6} + 500 \times \frac{1}{6} = 116,6$$

¿qué significa? es el valor medio a la larga: si se juega un número grande de veces la ganancia final es como si en cada jugada se hubiera ganado 116,6 pts. Si la apuesta costara menos de eso el juego sería ventajoso para el jugador (así se enriqueció Voltaire), si costara más, para la banca. (llamar a ésto honestidad del juego le costó el puesto de ministro a Laplace).

Varianza:

Se define como:

$$\sigma_X^2 = E\left[\left(X - \mu_X\right)^2\right]$$

aunque para el cálculo se suele usar esta otra fórmula equivalente:

$$\sigma_X^2 = E\left[X^2\right] - \mu_X{}^2$$

¿Qué mide la varianza? Mide la dispersión de la variable alrededor de la media.

3.2.Distribuciones discretas de probabilidad

Una distribución de probabilidad es una representación de todos los resultados posibles de algún experimento y de la probabilidad relacionada con cada uno.

Una distribución de probabilidad es discreta cuando los resultados posibles del experimento son obtenidos de variables aleatorias discretas, es decir, de variables que sólo puede tomar

ciertos valores, con frecuencia números enteros, y que resultan principalmente del proceso de conteo.

Ejemplos de variables aleatorias discretas son:

- ✓ Número de caras al lanzar una moneda
- ✓ El resultado del lanzamiento de un dado
- ✓ Número de hijos de una familia
- ✓ Número de estudiantes de una universidad

Ejemplo ilustrativo: Sea el experimento aleatorio de lanzar 2 monedas al aire. Determinar la distribución de probabilidades del número de caras.

El espacio muestral es S = {CC, CS, SC, SS}

La probabilidad de cada punto muestral es de 1/4, es decir, P(CC) = P(CS) = P(SC) = P(SS) = 1/4

La distribución de probabilidades del número de caras se presenta en el siguiente diagrama:

Interpretación:

- ✓ La probabilidad de obtener 0 caras al lanzar 2 monedas al aire es de 1/4 = 0,25 = 25%
- ✓ La probabilidad de obtener una cara al lanzar 2 monedas al aire es de 2/4 = 0,5 = 50%
- ✓ La probabilidad de obtener 2 caras al lanzar 2 monedas al aire es de 1/4 = 0,25 = 25%

3.2.1. La media y la varianza de las distribuciones discretas

A. Media

La media llamada también valor esperado, esperanza matemática o simplemente esperanza de una distribución de probabilidad discreta es la media aritmética ponderada de todos los resultados posibles en los cuales los pesos son las probabilidades respectivas de tales resultados. Se halla multiplicando cada resultado posible por su probabilidad y sumando los resultados. Se expresa mediante la siguiente fórmula:

$$\mu = \mathrm{E}(\mathrm{X}) = \Sigma(x_i \cdot P(x_i))$$

Donde:

- ✓ $\mu = \mathrm{E}(\mathrm{X}) =$ Media, Valor Esperado, Esperanza Matemática o simplemente Esperanza
- ✓ x_i = Posible resultado
- ✓ $P(x_i) =$ Probabilidad del posible resultado

B. Varianza

La varianza es el promedio de las desviaciones al cuadrado con respecto a la media. La varianza mide la dispersión de los resultados alrededor de la media y se halla calculando las diferencias entre cada uno de los resultados y su media, luego tales diferencias se elevan al cuadrado y se multiplican por sus respectivas probabilidades, y finalmente se suman los resultados. Se expresa mediante la siguiente fórmula:

$$\sigma^2 = \Sigma[(x_i - \mu)^2 \cdot P(x_i)]$$

Nota: La varianza se expresa en unidades al cuadrado, por lo que es necesario calcular la desviación estándar que se expresa en las mismas unidades que la variable aleatoria y que por lo tanto tiene una interpretación más lógica de la dispersión de los resultados alrededor de la media. La desviación estándar se calcula así: $\sigma = \sqrt{\sigma^2}$

C. Características:

1. Es generada por una variable discreta (x).

- ✓ x→Variable que solo toma valores enteros
- ✓ x→0, 1, 2, 3, 4, 5, 6, 7, 8, ... etc,etc.

2. p(xi)≥0 Las probabilidades asociadas a cada uno de los valores que toma x deben ser mayores o iguales a cero.

3.Σp(xi) = 1 La sumatoria de las probabilidades asociadas a cada uno de los valores que toma x debe ser igual a 1.

3.3.Distribuciones continuas de probabilidad

3.3.1. Características:

1. Es generada por una variable continua (x).

- ✓ x→ Es una variable que puede tomar tanto valores enteros como fraccionarios.
- ✓ x→ 1.0, 3.7, 4.0, 4.6, 7.9, 8.0, 8.3, 11.5, ...,∞

2. f(x)≥0 Las probabilidades asociadas a cada uno de los valores que toma x deben ser mayores o iguales a cero. Dicho de otra forma, la función de densidad de probabilidad deberá tomar solo valores mayores o iguales a cero. La función de densidad de probabilidad sólo puede estar definida en los cuadrantes I y II.

3. $\int_{-\infty}^{\infty} f(x)dx = 1$ La sumatoria de las probabilidades asociadas a cada uno de los valores que toma x debe ser igual a 1. El área definida bajo la función de densidad de probabilidad deberá ser de 1.

3.3.2. Calculo de media y desviación estándar para una distribución continua

- ✓ **Media o valor esperado de x**.- Para calcular la media de una distribución de probabilidad continua se utiliza la siguiente fórmula:

$$\mu = \int_{-\infty}^{\infty} xf(x)dx$$

Donde:

μ = E(x) = media o valor esperado de la distribución

x = variable aleatoria continua

f(x) = función de densidad de la distribución de probabilidad

- ✓ **Desviación estándar.** - La fórmula para determinar la desviación estándar de una distribución continua es;

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 * f(x)dx$$

luego:

$$\sigma = \sqrt{\sigma^2}$$

Ejemplo:

Para la siguiente función,

$f(x) = \frac{1}{9}x^2$ cuando $0 \leq x \leq 3$, f(x) = 0 para cualquier otro valor

a) Diga si esta función nos define una distribución de probabilidad.

b) Si la función define una distribución de probabilidad, entonces, determine su media y desviación estándar.

c) Determine la probabilidad de que $1 \leq x < 2$.

Solución:

a) Para verificar que la función nos define una distribución de probabilidad, es necesario que cumpla con las características que se habían mencionado.

- ✓ $x \rightarrow$ sí es una variable continua porque puede tomar cualquier valor entre 0 y 3
- ✓ $f(x) \geq 0$, lo que se comprueba si damos diferentes valores a x para ver que valores toma f(x), dándonos cuenta de que efectivamente f(x) solo toma valores mayores o iguales a cero.

x	f(x)
0	0.0
0.5	0.02778
1.0	0.11111
1.4	0.21778
2.1	0.49
2.7	0.81
3.0	1.0

- ✓ Para comprobar que la sumatoria de las probabilidades que toma cada valor de x es de 1, se integra la función de 0 a 3 como se muestra a continuación:

$$A = \int_{-\infty}^{\infty} f(x)dx = \int_{0}^{3} \frac{1}{9}x^{2}\, dx = \frac{1}{9}\left(\frac{x^{2+1}}{2+1}\right) \Big| = \frac{1}{27}(3^{3} - 0^{3}) = \frac{1}{27}(27-0) = 1$$

A= área bajo la función

Con las operaciones anteriores comprobamos que la función $\frac{1}{9}x^{2}$ sí nos define una distribución de probabilidad continua.

b) Cálculo de media y desviación estándar.

$$\mu = \int_{-\infty}^{\infty} x * f(x)dx = \int_{0}^{3} x(\tfrac{1}{9}x^{2})dx = \int_{0}^{3} \frac{1}{9}x^{3}\, dx = \frac{1}{9}\left(\frac{x^{4}}{4}\right) \Big| =$$

$$= \frac{1}{36}(3^{4} - 0^{4}) = \frac{1}{36}(81-0) = \frac{81}{36} = 2.25$$

$$\sigma^{2} = \int_{-\infty}^{\infty} (x-\mu)^{2} * f(x)dx = \int_{0}^{3} (x-2.25)^{2} * \left(\frac{1}{9}x^{2}\right)dx =$$

$$= \int_{0}^{3} (x^{2} - 4.5x + 5.0625)\left(\frac{1}{9}x^{2}\right)dx = \int_{0}^{3} \left(\frac{x^{4}}{9} - \frac{x^{3}}{2} + \frac{5.0625x^{2}}{9}\right)dx =$$

$$= \frac{x^{5}}{45} \Big| - \frac{x^{4}}{8} \Big| + \frac{5.0625x^{3}}{27} \Big| = \frac{(3)^{5}}{45} - \frac{(3)^{4}}{8} + \frac{5.0625(3)^{3}}{27} =$$

$$= \frac{243}{45} - \frac{81}{8} + \frac{136.6875}{27} = 5.4 - 10.125 + 5.0625 = 0.3375$$

$$\sigma = \sqrt{\sigma^{2}} = \sqrt{0.3375} = 0.5809$$

Las barras nos indican la evaluación de la integral entre 0 y 3.

c) $$p(1 \le x < 2) = \int_{1}^{2} f(x)dx = \int_{1}^{2} \frac{1}{9}x^{2}\, dx = \frac{1}{9}\left(\frac{x^{3}}{3}\right) \Big| = \frac{1}{9}\left(\frac{2^{3}}{3} - \frac{0^{3}}{3}\right) = \frac{1}{9}\left(\frac{8}{3}\right) = \frac{8}{27} = 0.2963$$

La barra nos indica la evaluación de la integral de 1 a 2.

Con las operaciones anteriores nos damos cuenta que para evaluar probabilidades para variables de tipo continuo, es necesario evaluar la función de densidad de probabilidad en el rango de valores que se desea; que vendría siendo el área que se encuentra entre f(x) y el eje de las x y entre el rango de valores definidos por la variable x.

3.4.Distribuciones empíricas

Los percentiles empíricos se calculan a partir de la función de distribución empírica definida por los valores de la serie con la que se trabaja ordenada desde el valor menor al mayor, y asignando a cada valor ordenado su probabilidad calculada según la expresión:

$$\text{Prob } (X \leq xi) = i/(N+1).$$

Donde "i" representa el número de orden que ocupa el valor "x" en la serie de datos ordenada en orden creciente y "N" el número total de datos. La probabilidad correspondiente al 20, 40, 50, 60 ó 80 por ciento se obtienen por interpolación lineal, considerando las probabilidades asignadas a cada dato ordenado.

Ejercicio. Se pide calcular los valores de los percentiles 20 y 40 mediante la función de distribución empírica, de la siguiente serie de valores:

102.2 96.3 377.7 119.9

221.1 32 153.8 199

261.9 58.7 160 209.8

270 60.4 171.9 142

138.3 83.5 172.1 148.5

13.5 289.4 183.6 269.4

18.1 299.9 197.9

118 110.5 300.7

Solución.

Se deben ordenar los datos de precipitación en orden creciente, y asignar a cada valor de precipitación su probabilidad empírica en función del orden de situación del valor y del número de datos. Así para los dos primeros valores y para los dos últimos tendremos:

Nº orden	Precipitación	Probabilidad
1 (primero)	13.5 mm	Prob (X≤13.5) = i/(N +1) = 1/(30+1) = 0.03226 (3.226 %)
2 (segundo)	18.1 mm	Prob (X≤18.1) = i/(N +1) = 2/(30+1) = 0.06452 (6.452 %)
...		
29 (vigésimo nono)	300,7 mm	Prob (X≤300,7) = i/(N +1) = 29/(30+1) = 0.9354 (93,55 %)
30 (trigésimo)	377,7 mm	Prob (X≤377,7) = i/(N +1) = 30/(30+1) = 0.9677 (96,77 %)

Los valores de precipitación ordenados desde el menor al mayor para los treinta años de la serie y los valores de probabilidad asignados son:

Nºorden	**Prob %**	**Prec (mm)**
1	3.226	13.5
2	6.452	18.1

3	9.677	32
4	12.9	58.7
5	16.13	60.4
6	19.35	83.5
7	22.58	96.3
8	25.81	102.2
9	29.03	110.5
10	32.26	118
11	35.48	119.9
12	38.71	138.3
13	41.94	142
14	45.16	148.5
15	48.39	153.8
16	51.61	160
17	54.84	171.9
18	58.06	172.1
19	61.29	183.6
20	64.52	197.9
21	67.74	199
22	70.97	209.8
23	74.19	221.1
24	77.42	261.9
25	80.65	269.4
26	83.87	270
27	87.1	289.4
28	90.32	299.9
29	93.55	300.7
30	96.77	377.7

El percentil 20 se obtiene por interpolación sabiendo que será un volumen de precipitación entre el valor que está en la posición sexta (19,35 %) y el valor que está en la posición séptima (22,58 %)

P19.35 83.5 P20 86.07 P22,58 96.3

El percentil 40 se obtiene por interpolación sabiendo que será un volumen de precipitación entre el valor que está en la posición duodécima (38,71 %) y el valor que está en la posición decimotercera (41,94 %)

P38.71 138.3 P40 139.78 P41.94 142

3.5.Distribuciones de probabilidad conjunta

En probabilidad, dados dos eventos aleatorios X y Y, la distribución conjunta de X y Y es la distribución de probabilidad de la intersección de eventos de X y Y, esto es, de los eventos X e Y ocurriendo de forma simultánea. En el caso de solo dos variables aleatorias se denomina una distribución bivariada, pero el concepto se generaliza a cualquier número de eventos o variables aleatorias.

Caso discreto

Para variables aleatorias discretas, la función de probabilidad conjunta está dada por:

$$\begin{aligned}\mathrm{P}(X = x \text{ y } Y = y) &= \mathrm{P}(Y = y \mid X = x) \cdot \mathrm{P}(X = x) \\ &= \mathrm{P}(X = x \mid Y = y) \cdot \mathrm{P}(Y = y).\end{aligned}$$

Dadas esas probabilidades, se tiene que:

$$\sum_x \sum_y \mathrm{P}(X = x \text{ y } Y = y) = 1.$$

Caso continuo

Para las variables aleatorias continuas la función de densidad de probabilidad conjunta puede ser escrita como fX,Y(x, y) teniendo:

$$f_{X,Y}(x, y) <= f_{Y|X}(y|x) f_X(x) = f_{X|Y}(x|y) f_Y(y)$$

Donde fY|X(y|x) y fX|Y(x|y) dan la Probabilidad condicionada de Y dado X = x y de X dado Y = y respectivamente, y fX(x) y fY(y) dada la distribución marginal para X y Y respectivamente.

De nuevo, dado que son distribuciones de probabilidad:

$$\int_x \int_y f_{X,Y}(x, y)\, dy\, dx = 1.$$

4. ESPERANZA MATEMÁTICA

4.1. Media de una variable aleatoria

4.1.1. Variables aleatorias discretas

Una variable aleatoria se dirá discreta si el conjunto de valores que toma es un conjunto numerable, es decir, que solo puede tomar unos valores concretos. Dicho conjunto lo denotaremos por: {x1, x2, x3,...., xk}

Toda variable aleatoria discreta tiene asociada una función de probabilidad, que a cada valor, le marca la probabilidad de que la variable tome dicho valor. Esta probabilidad viene a jugar el mismo papel que la frecuencia relativa en los temas de estadística.

4.1.2. Parámetros de una variable aleatoria

La ventaja de trabajar con variables aleatorias es que podemos hacer cálculos que adquieren signifcado sobre el comportamiento de la variable. En una variable aleatoria, podemos calcular **todos** los parámetros que habíamos visto en la estadística unidimensional: media, varianza moda, mediana, percentiles, desviaciones, etc, aunque nosotros vamos a centrarnos en las dos primeras, la media y la varianza, (bueno o la desviación típica que era la raíz de la varianza si recuerdas)

MEDIA**:** La media de una variable aleatoria se llama ESPERANZA MATEMÁTICA, se representa por E(X) o por μ y viene a darnos el ***"valor esperado"*** **de la variable al realizar el experimento aleatorio.** La fórmula para calcularla es

$$E(X) = \sum_{i=1}^{k} x_i \cdot p_i = x_1 \cdot p_1 + x_2 \cdot p_2 + \ldots\ldots + x_k \cdot p_k$$

VARIANZA: El significado es el mismo que en la estadística. Aporta una medida sobre la dispersión de los valores de X. Para calcularla usamos una de las dos fórmulas, aunque es más aconsejable la segunda:

$$VAR(X) = \sum_{i=1}^{k} (x_i - \mu)^2 \cdot p_i \qquad VAR(X) = \sum_{i=1}^{k} x_i^2 \cdot p_i - \mu^2$$

Veamos un par de ejemplo:

Calcula la media y la varianza de la variable del Ejemplo

xi	pi = P[X = xi]
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

$$E(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{21}{6} = 3.5$$

Lo que significa que el valor esperado en el lanzamiento de un dado es 3,5

$$Var(X) = \left(1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + 3^2 \cdot \frac{1}{6} + 4^2 \cdot \frac{1}{6} + 5^2 \cdot \frac{1}{6} + 6^2 \cdot \frac{1}{6} \right) - 3.5^2 = \frac{91}{6} - 12.25 = 2.916$$

y si queremos calcular la desviación típica, haríamos la raíz cuadrada de ese resultado y obtendríamos, 1.708

4.2. Varianza y covarianza

3.5.1. La Varianza

Otra forma para asegurar que las diferencias entre la media y los puntos de un valor positivo, es elevándola al cuadrado. Al promedio de estas distancias al cuadrado se le conoce como varianza.

Varianza (S^2 o σ^2): Es el resultado de la división de la sumatoria de las distancias existentes entre cada dato y su media aritmética elevadas al cuadrado, y el número total de datos.

Distinguimos dos símbolos para identificar la varianza: S^2 para datos muestrales, y σ^2 para datos poblacionales. Note que la fórmula para la varianza muestral presenta en su denominador al tamaño de la muestra menos uno, tendencia adoptada por los estadísticos para denotar una varianza más conservadora.

Al igual que ocurre con la desviación media, podemos definir las fórmulas para datos agrupados en tablas tipo A y tipo B. Para las tablas tipo A tenemos:

Una advertencia en el uso de esta medida, es que al elevar las distancias al cuadrado, automáticamente se elevan las unidades. Por ejemplo, si unidad trabajada en los datos es centímetros, la varianza da como resultados centímetros al cuadrado.

Ejemplo: Varianza para datos no agrupados

La siguiente muestra representa las edades de 25 personas sometidas a un análisis de preferencias para un estudio de mercado.

25	19	21	35	44
20	27	32	38	33
18	30	19	29	33
26	24	28	39	31
31	18	17	30	27

Determinar la varianza.

Solución:

PASO 1: Calcular la media aritmética.

PASO 2: Calcular la varianza

En este punto, la varianza es identificada por S^2.

La varianza equivale a 51,8567. Por elevar las unidades al cuadrado, carece de un significado contextual dentro del análisis descriptivo del caso.

Ejemplo: Varianza para datos agrupados

Calcular la varianza a partir de la siguiente tabla de frecuencia (suponga que los datos son poblacionales).

Ni	Lm	Ls	f	Mc
1	[15	17)	2	16
2	[17	19)	5	18
3	[19	21)	13	20
4	[21	23)	4	22
5	[23	25]	1	24
Total		**25**		

Solución:

PASO 1: Calcular la media aritmética.

PASO 2: Calcular la varianza

En este punto, la varianza es identificada por S^2.

Cálculo de la varianza en Excel

Excel posee dos funciones propias para el cálculo de la media, diferenciando los datos muestrales de los datos poblacionales.

VAR: Calcula la varianza de una muestra.

Formato: VAR(número1;número2;…)

Categoría: Estadísticas

VARP: Calcula la varianza de todos los datos de una población.

Formato: VARP(número1;número2;…)

Categoría: Estadísticas

Mostremos su funcionamiento calculando la varianza en ambos casos a partir de los siguientes datos:

138,2	195,8	124,5	101,7	137,1	130,3
110,0	101,4	104,5	128,5	135,5	197,5
159,6	140,7	103,2	134,3	191	180,6
189,9	186,3	116,4	155,3	146,6	199,1
188,4	113,8	121,9	135,7	142,6	125,6

Los datos copiados en Excel desde la celda B2 deberían verse como sigue:

Si los datos provienen de una muestra, emplearemos la función **VAR**, en cuyo denominador se tendría el valor 29 en vez de 30, equivalente al tamaño de la muestra. Activemos esta función en la celda B8.

El resultado de la varianza muestral es de 1034,138051.

En la celda B9 calculemos la varianza para datos poblacionales.

La función de la varianza **VARP**, divide la sumatoria de las distancias al cuadrado por los 30 datos, dando como resultado un valor menor que con la función **VAR** (la varianza para la muestra es un valor más conservador).

Para el cálculo de la varianza en datos agrupados en Excel, tomaremos la tabla de frecuencia dada en el ejemplo 5.2.2.

Calculemos la media en la celda B10.

En una columna adicional colocaremos las diferencias entre la marca de clase y la media elevadas al cuadrado multiplicadas por su frecuencia.

Analicemos la fórmula empleada desde la celda C3.

La celda B10 esta fija indicando la media aritmética. Aparece el operador ∧, la cual eleva al cuadrado lo que esta dentro del paréntesis. Esta distancia se multiplica por el número de veces que se repite (por su frecuencia). Al final calculamos su sumatoria.

En la celda B11 calculamos la varianza.

3.5.2. Covarianza

La **covarianza** de una variable bidimensional es la media aritmética de los productos de las desviaciones de cada una de las variables respecto a sus medias respectivas.

La **covarianza** se representa por $\mathbf{s_{xy}}$ o $\boldsymbol{\sigma_{xy}}$.

$$\sigma_{xy} = \frac{\sum f_i(x_i - \bar{x})(y_i - \bar{y})}{N}$$

$$\sigma_{xy} = \frac{\sum f_i x_i y_i}{N} - \bar{x}\bar{y}$$

La **covarianza** indica el sentido de la correlación entre las variables

Si $\boldsymbol{\sigma_{xy} > 0}$ la correlación es directa.

Si $\boldsymbol{\sigma_{xy} < 0}$ la correlación es inversa.

La covarianza presenta como inconveniente, el hecho de que su valor depende de la escala elegida para los ejes.

Es decir, la **covarianza** variará si expresamos la altura en metros o en centímetros. También variará si el dinero lo expresamos en euros o en dólares.

Ejemplos

Las notas de 12 alumnos de una clase en Matemáticas y Física son las siguientes:

Matemáticas	Física
2	1
3	3
4	2
4	4
5	4
6	4
6	6
7	4
7	6
8	7
10	9
10	10

Hallar la **covarianza** de la distribución.

xi	yi	xi · yi
2	1	2
3	3	9
4	2	8
4	4	16
5	4	20
6	4	24
6	6	36
7	4	28
7	6	42
8	7	56
10	9	90
10	10	100
72	60	431

Después de tabular los datos hallamos las medias aritméticas:

$$\bar{x} = \frac{72}{12} = 6 \qquad \bar{y} = \frac{60}{12} = 5$$

$$\sigma_{xy} = \frac{431}{12} - 6 \cdot 5 = 5.92$$

Los valores de dos variables X e Y se distribuyen según la tabla siguiente:

Y/X	0	2	4
1	2	1	3
2	1	4	2
3	2	5	0

Hallar la **covarianza** de la distribución.

En primer lugar, convertimos la tabla de doble entrada en tabla simple y calculamos las medias aritméticas.

xi	yi	fi	xi · fi	yi · fi	xi · yi · fi
0	1	2	0	2	0
0	2	1	0	2	0
0	3	2	0	6	0
2	1	1	2	1	2
2	2	4	8	8	16
2	3	5	10	15	30
4	1	3	12	3	12
4	2	2	8	4	16
		20	40	41	76

$$\bar{x} = \frac{40}{20} = 2 \qquad \bar{y} = \frac{41}{20} = 2.05$$

$$\sigma_{xy} = \frac{76}{20} - 2 \cdot 2.05 = -0.3$$

4.3. Medias y varianzas de combinaciones lineales de variables aleatorias

Dadas las variables aleatorias X1, X2, . . . , Xn y las constantes c1, c2, . . . , cn, definimos la siguiente combinación lineal: Y = c1X1 + c2X2 + . . . cnXn.

Entonces tenemos que:

$$E[Y] = E[\sum_{i=1}^{n} c_i X_i] = \sum_{i=1}^{n} c_i E[X_i]$$

El cálculo de σ_Y^2 es mas complejo. De la definición de covarianza, es fácil verificar que ´esta cumple con las siguientes propiedades:

- $Cov(X, X) = Var(X)$
- $Cov(aX, Y) = a \cdot Cov(X, Y)$
- $Cov(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_i) = \sum_{i=1}^{n} \sum_{j=1}^{m} Cov(Y_j, X_i)$

De lo que se desprende como corolario que:

$$\sigma_Y^2 = Var\left(\sum_{i=1}^{n} c_i X_i\right) = \sum_{i=1}^{n} c_i^2 Var(X_i) + \sum_{i=1}^{n} \sum_{j \neq i}^{n} {}_{j=1} c_i^2 c_j^2 Cov(X_j, X_i)$$

En el caso de que $X_1 X_2, \ldots, X_n$ sean independientes se cumplirá que $\sigma_Y^2 = \sum_{i=1}^{n} c_i \sigma_{X_i}^2$.

El caso de querer encontrar la distribución de la combinación lineal de variables aleatorias tendremos que emplear la convolución. Sean X1, X2 dos variables aleatorias discretas, y sea Y = X1 + X2. En tal caso

$$P(Y = y) = \sum_{i=0}^{y} P(X_1 = i)P(X_2 = y - i)$$

4.4. Teorema de Chebyshev

4.4.1. Desigualdad de Chebyshev

La desigualdad de Chebyshev es un teorema utilizado en estadística que proporciona una estimación conservadora (intervalo de confianza) de la probabilidad de que una variable aleatoria con varianza finita, se sitúe a una cierta distancia de su esperanza matemática o de su media.

Su expresión formal es la siguiente:

$$P(|X - \mu| > k6) \leq \frac{1}{k^2}$$

X = Valor estimado

μ = Esperanza matemática del valor estimado

б = Desviación típica del valor esperado

k = Número de desviaciones típicas

Partiendo de esta expresión general y desarrollando la parte que queda dentro del valor absoluto tendríamos lo siguiente:

$$P(\mu - k6 < X < \mu + k6) \geq 1 - \frac{1}{k^2}$$

Si prestamos atención a la expresión anterior, se aprecia que la parte de la izquierda no es más es un intervalo de confianza. Este nos ofrece tanto una cota inferior, como una superior para el valor estimado. Por lo tanto, la desigualdad de Chebyshev nos dice la probabilidad mínima, de que el parámetro poblacional se encuentre dentro de una determinada cantidad de desviaciones típicas por encima o por debajo de su media. O dicho de otra manera, nos da la probabilidad de que de que el parámetro poblacional se encuentre dentro de ese intervalo de confianza.

La desigualdad de Chebyshev proporciona cotas aproximadas para el valor estimado. A pesar de tener cierto grado de imprecisión, es un teorema bastante útil dado que se puede aplicar a un amplio abanico de variables aleatorias independientemente de sus distribuciones. La única restricción para poder utilizar esta desigualdad es que k tiene que ser mayor que 1 (k>1).

Ejemplo de aplicación de la desigualdad de Chebyshev

Supongamos que somos gestores de un fondo de inversión. La cartera que estamos gestionando tiene una rentabilidad media del 8,14% y una desviación típica del 5,12%. Para

saber por ejemplo que porcentaje de nuestros retornos se encuentran al menos a 3 desviaciones típicas de nuestra rentabilidad media simplemente aplicaríamos la formula anterior de la expresión 2.

k = 1,96

Sustituyendo el valor de k: 1-1/1,96^2 = 0,739 = 73,9%

Esto quiere decir que hay un 73,9% de los resultados que están en el intervalo de confianza situado a 1,96 desviaciones típcas de la media.

Realicemos el ejemplo anterior para valores distintos de k.

k = 2,46

k = 3

Sustituyendo el valor de k: 1-1/2,46^2 = 0,835 = 83,5%

Sustituyendo el valor de k: 1-1/3^2 = 0,889 = 88,9%

Hay un 83,5% de los datos que están a una distancia de 2,46 desviaciones típicas de la media y un 88,9% que están a 3 desviaciones típicas de la media.

Utilizando la desigualdad de Chebyshev, es sencillo deducir que a mayor valor de K (mayor desviación del valor estimado sobre su media) mayor probabilidad de que la variable aleatoria se encuentro dentro del intervalo acotado.

5. DISTRIBUCIÓN DE PROBABILIDAD DISCRETA

5.1. conceptos Generales

Uno de los objetivos de la estadística es el conocimiento cuantitativo de una determinada

parcela de la realidad. Para ello, es necesario construir un modelo de esta realidad particular objeto de estudio, partiendo de la premisa de que lo real es siempre más complejo y multiforme que cualquier modelo que se pueda construir. De todas formas, la formulación de modelos aceptados por las instituciones responsables y por los usuarios, permite obviar la existencia del error o distancia entre la realidad y el modelo.

Los modelos teóricos a los que se hace referencia se reducen en muchos casos a (o incluyen

en su formulación) funciones de probabilidad. La teoría de la probabilidad tiene su origen en el estudio de los juegos de azar, que impulsaron los primeros estudios sobre cálculo de probabilidades en el siglo XVI, aunque no es hasta el siglo XVIII cuando se aborda la probabilidad desde una perspectiva matemática con la demostración de la “ley débil de los grandes números” según la cual, al aumentar el número de pruebas, la frecuencia de un suceso tiende a aproximarse a un número fijo denominado probabilidad. Este enfoque, denominado enfoque frecuentista, se modela matemáticamente en el siglo XX cuando Kolmogorov formula la teoría axiomática de la probabilidad1. Dicha teoría define la probabilidad como una función que asigna a cada posible resultado de un experimento aleatorio un valor no negativo, de forma que se cumpla la propiedad aditiva. La definición axiomática establece las reglas que deben cumplir las probabilidades, aunque no asigna valores concretos.

Uno de los conceptos más importantes de la teoría de probabilidades es el de variable

aleatoria que, intuitivamente, puede definirse como cualquier característica medible que toma diferentes valores con probabilidades determinadas. Toda variable aleatoria posee una

distribución de probabilidad que describe su comportamiento (vale decir, que desagrega el 1 a lo largo de los valores posibles de la variable). Si la variable es discreta, es decir, si toma valores aislados dentro de un intervalo, su distribución de probabilidad especifica todos los valores posibles de la variable junto con la probabilidad de que cada uno ocurra. En el caso continuo, es decir, cuando la variable puede tomar cualquier valor de un intervalo, la distribución de probabilidad permite determinar las probabilidades correspondientes a con subintervalos de valores. Una forma usual de describir la distribución de probabilidad de una variable aleatoria es mediante la denominada función de densidad, en tanto que lo que se conoce como función de distribución representa las probabilidades acumuladas2-7.

Una de las preocupaciones de los científicos ha sido construir modelos de distribuciones de

probabilidad que pudieran representar el comportamiento teórico de diferentes fenómenos aleatorios que aparecían en el mundo real. La pretensión de modelar lo observable ha constituido siempre una necesidad básica para el científico empírico, dado que a través de esas construcciones teóricas, los modelos, podía experimentar sobre aquello que la realidad no le permitía. Por otra parte, un modelo resulta extremadamente útil, siempre que se corresponda con la realidad que pretende representar o predecir, de manera que ponga de relieve las propiedades más importantes del mundo que nos rodea, aunque sea a costa de la simplificación que implica todo modelo.

En la práctica hay unas cuantas leyes de probabilidad teóricas, como son, por ejemplo, la ley binomial o la de Poisson para variables discretas o la ley normal para variables continuas, que sirven de modelo para representar las distribuciones empíricas más frecuentes

Así, por ejemplo, la variable “talla de un recién nacido” puede tener valores entre 47 cm y 53

cm, pero no todos los valores tienen la misma probabilidad, porque las más frecuentes son las tallas próximas a los 50 cm. En este caso la ley normal se adapta satisfactoriamente a la distribución de probabilidad empírica, que se obtendría con una muestra grande de casos.

Se ofrece, en este módulo, procedimientos usuales para calcular probabilidades y sus

inversas, para un conjunto bastante amplio de funciones de distribución, discretas y continuas, que son habituales en el proceso de modelación. Por ejemplo, el conjunto de distribuciones pertenecientes a la familia exponencial es de uso frecuente en metodologías como el análisis de supervivencia o el Modelo Lineal Generalizado. Otras distribuciones son comunes y habituales en el campo de actuación de disciplinas tales como la economía, la biología, etc.

Cuando la opción elegida es el cálculo de una probabilidad dado un punto x de la

distribución, se presentan en todos los casos dos resultados: la probabilidad acumulada hasta ese punto, o la probabilidad de que la variable tome valores inferiores o iguales a x (cola izquierda) y la probabilidad de valores superiores a x (cola derecha). En el caso continuo, la probabilidad de que la variable sea igual a cualquier punto es igual a cero; por tanto, no influye en las colas el hecho de incluir o excluir el punto x. Hay un tercer resultado que el programa presenta sólo para las distribuciones continuas simétricas (normal, logística y t de Student): la probabilidad de dos colas, es decir, la probabilidad que queda a ambos lados del intervalo (-x, x) ó (x, -x), según el punto sea positivo o negativo, respectivamente.

Asimismo, los resultados de Epidat incluyen la media y la varianza de la correspondiente distribución, así como la mediana y/o la moda en el caso de las distribuciones continuas. Epidat

también ofrece la posibilidad de representar, gráficamente, las funciones de distribución y densidad.

5.2. distribución uniforme discreta

Las distribuciones discretas incluidas en el módulo de “Cálculo de probabilidades” son:

- ✓ Distribución Uniforme discreta
- ✓ Geométrica
- ✓ Binomial
- ✓ Binomial Negativa
- ✓ Hipergeométrica
- ✓ Poisson

5.2.1. Distribución Uniforme discreta (a,b)

Describe el comportamiento de una variable discreta que puede tomar n valores distintos

con la misma probabilidad cada uno de ellos. Un caso particular de esta distribución, que es la que se incluye en este módulo de Epidat , ocurre cuando los valores son enteros consecutivos. Esta distribución asigna igual probabilidad a todos los valores enteros entre el límite inferior y el límite superior que definen el recorrido de la variable. Si la variable puede tomar valores entre a y b, debe ocurrir que b sea mayor que a, y la variable toma los valores enteros empezando por a, a+1, a+2, etc. hasta el valor máximo b. Por ejemplo, cuando se observa el número obtenido tras el lanzamiento de un dado perfecto, los valores posibles

siguen una distribución uniforme discreta en {1, 2, 3, 4, 5, 6}, y la probabilidad de cada cara es 1/6

Valores:

x: a, a+1, a+2, ..., b, números enteros

Parámetros:

a: mínimo, a entero

b: máximo, b entero con $a < b$

Ejercicio

El temario de un examen para un proceso selectivo contiene 50 temas, de los cuales se elegirá

uno por sorteo. Si una persona no ha estudiado los 15 últimos temas ¿Cuál es la probabilidad de que apruebe el examen?

La variable que representa el número del tema seleccionado para el examen sigue una

distribución uniforme con parámetros a=1 y b=50. La persona aprueba el examen si le toca un tema del 1 al 35; por tanto, la probabilidad que se pide es la cola a la izquierda de 35. Para obtener los resultados en Epidat 3.1 basta con proporcionarle los parámetros de la distribución, y seleccionar calcular probabilidades para el punto 35.

Resultados con Epidat 3.1

Resultados con Epidat 3.1

```
Cálculo de probabilidades. Distribuciones discretas

Uniforme discreta (a,b)
a : Mínimo                          1
b : Máximo                         50
Punto K                            35

Probabilidad    Pr[X=k]          0,0200
Cola Izquierda  Pr[X<=k]         0,7000
Cola Derecha    Pr[X>k]          0,3000

Media                           25,5000
Varianza                       208,2500
```

La persona tiene una probabilidad de aprobar igual a 0,7.

5.2.2. Distribución Binomial (n,p)

La distribución binomial es una distribución discreta muy importante que surge en muchas aplicaciones bioestadísticas.

Esta distribución aparece de forma natural al realizar repeticiones independientes de un experimento que tenga respuesta binaria, generalmente clasificada como "éxito" o "fracaso". Por ejemplo, esa respuesta puede ser el hábito de fumar (sí/no), si un paciente hospitalizado desarrolla o no una infección, o si un artículo de un lote es o no defectuoso. La variable discreta que cuenta el número de éxitos en n pruebas independientes de ese experimento, cada una de ellas con la misma probabilidad de "éxito" igual a p, sigue una distribución binomial de parámetros n y p. Este modelo se aplica a poblaciones finitas de las que se toma elementos al azar con reemplazo, y también a poblaciones conceptualmente infinitas, como por ejemplo las piezas que produce una máquina, siempre que el proceso de producción sea estable (la proporción de piezas defectuosas se mantiene constante a largo plazo) y sin memoria (el resultado de cada pieza no depende de las anteriores).

Un ejemplo de variable binomial puede ser el número de pacientes ingresados en una unidad hospitalaria que desarrollan una infección nosocomial

Un caso particular se tiene cuando n=1, que da lugar a la distribución de Bernoulli.

Valores:

x: 0, 1, 2, ..., n

Parámetros:

n: número de pruebas, $n > 0$ entero p: probabilidad de éxito, $0 < p < 1$

Ejercicio

En un examen formado por 20 preguntas, cada una de las cuales se responde declarando "verdadero" o "falso", el alumno sabe que, históricamente, en el 75% de los casos la respuesta correcta es "verdadero" y decide responder al examen tirando dos monedas, pone "falso" si ambas monedas muestran una cara y "verdadero" si al menos hay una cruz. Se desea saber qué probabilidad hay de que tenga al menos 14 aciertos.

Hay que proporcionarle a Epidat 3.1 los parámetros de la distribución y el punto k a partir del cual se calculará la probabilidad. En este caso n=20, p=0,75 y el punto k=14.

Resultados con Epidat 3.1

Resultados con Epidat 3.1

```
Cálculo de probabilidades. Distribuciones discretas

Binomial (n,p)
n: Número de pruebas                        20
p: Probabilidad de éxito          0,7500
Punto K                              14

Probabilidad    Pr[X=k]           0,1686
Cola Izquierda  Pr[X<=k]          0,3828
Cola Derecha    Pr[X>k]           0,6172

Media                            15,0000
Varianza                          3,7500
```

La probabilidad de que el alumno tenga más de 14 aciertos se sitúa en 0,61.

5.2.3. Distribución Hipergeométrica (N,R,n)

La distribución hipergeométrica suele aparecer en procesos muestrales sin reemplazo, en los que se investiga la presencia o ausencia de cierta característica. Piénsese, por ejemplo, en un procedimiento de control de calidad en una empresa farmacéutica, durante el cual se extraen muestras de las cápsulas fabricadas y se someten a análisis para determinar su composición. Durante las pruebas, las cápsulas son destruidas y no pueden ser devueltas al lote del que provienen. En esta situación, la variable que cuenta el número de cápsulas que no cumplen los criterios de calidad establecidos sigue una distribución hipergeométrica. Por tanto, esta distribución es la equivalente a la binomial, pero cuando el muestreo se hace sin reemplazo.

Esta distribución se puede ilustrar del modo siguiente: se tiene una población finita con N elementos, de los cuales R tienen una determinada característica que se llama "éxito" (diabetes, obesidad, hábito de fumar, etc.). El número de "éxitos" en una muestra aleatoria de tamaño n, extraída sin reemplazo de la población, es una variable aleatoria con distribución hipergeométrica de parámetros N, R y n

Cuando el tamaño de la población es grande, los muestreos con y sin reemplazo son equivalentes, por lo que la distribución hipergeométrica se aproxima en tal caso a la binomial.

Valores:

x: $max\{0,n-(N-R)\}, ..., min\{R,n\}$, donde $max\{0,n-(N-R)\}$ indica el valor máximo entre 0 y $n-(N-R)$ y $min\{R,n\}$ indica el valor mínimo entre R y n.

Parámetros:

N: tamaño de la población, $N>0$ entero

R: número de éxitos en la población, $R\geq 0$ entero n: número de pruebas, $n>0$ entero

Ejercicio

Se sabe que el 7% de los útiles quirúrgicos en un lote de 100 no cumplen ciertas especificaciones de calidad. Tomada una muestra al azar de 10 unidades sin reemplazo, interesa conocer la probabilidad de que no más de dos sean defectuosos.

El número de útiles defectuosos en el lote es R=0,07×100=7. Para un tamaño muestral de n=10, la probabilidad buscada es P{número de defectuosos ≤ 2}. Resultados con Epidat 3.1

Resultados con Epidat 3.1

```
Cálculo de probabilidades. Distribuciones discretas

Hipergeométrica (N,R,n)
N : Tamaño de la población          100
R : Número éxitos en la pob.          7
n : Número de pruebas                10
Punto K                               2

Probabilidad    Pr[X=k]           0,1235
Cola Izquierda  Pr[X<=k]          0,9792
Cola Derecha    Pr[X>k]           0,0208

Media                             0,7000
Varianza                          0,5918
```

La probabilidad de que a lo sumo haya dos útiles defectuosos en el lote es aproximadamente 0,98.

5.2.4. Distribución Geométrica (p)

Supóngase, que se efectúa repetidamente un experimento o prueba, que las repeticiones son independientes y que se está interesado en la ocurrencia o no de un suceso al que se refiere como "éxito", siendo la probabilidad de este suceso *p*. La distribución geométrica permite calcular la probabilidad de que tenga que realizarse un número k de repeticiones hasta obtener un éxito por primera vez. Así pues, se diferencia de la distribución binomial en que el número de repeticiones no está predeterminado, sino que es la variable aleatoria que se mide y, por otra parte, el conjunto de valores posibles de la variable es ilimitado.

Para ilustrar el empleo de esta distribución, se supone que cierto medicamento opera exitosamente ante la enfermedad para la cual fue concebido en el 80% de los casos a los que se aplica; la variable aleatoria "intentos fallidos en la aplicación del medicamento antes del primer éxito" sigue una distribución geométrica de parámetro p=0,8. Otro ejemplo de variable geométrica es el número de hijos hasta el nacimiento de la primera niña

La distribución geométrica se utiliza en la distribución de tiempos de espera, de manera que si los ensayos se realizan a intervalos regulares de tiempo, esta variable aleatoria proporciona el tiempo transcurrido hasta el primer éxito.

Esta distribución presenta la denominada "propiedad de Harkov" o de falta de memoria, que implica que la probabilidad de tener que esperar un tiempo *t* no depende del tiempo que ya haya transcurrido.

Valores:

x: 0, 1, 2, ...

Parámetros:

p: probabilidad de éxito, $0<p<1$

Ejercicio

La probabilidad de que cierto examen médico dé lugar a una reacción "positiva" es igual a 0,8, ¿cuál es la probabilidad de que ocurran menos de 5 reacciones "negativas" antes de la primera positiva?

La variable aleatoria "número de reacciones negativas antes de la primera positiva" sigue una distribución Geométrica con parámetro p=0,8.

Resultados con Epidat 3.1

```
Cálculo de probabilidades. Distribuciones discretas

Geométrica (p)
p : Probabilidad de éxito        0,8000
Punto K                               4

Probabilidad    Pr[X=k]          0,0013
Cola Izquierda  Pr[X<=k]         0,9997
Cola Derecha    Pr[X>k]          0,0003

Media                            0,2500
Varianza                         0,3125
```

La probabilidad de que ocurran menos de 5 reacciones "negativas" antes de la primera positiva es casi 1 (0,9997).

5.2.5. Distribución Binomial negativa (r,p)

Una generalización obvia de la distribución geométrica aparece si se supone que un experimento se continúa hasta que un determinado suceso, de probabilidad p, ocurre por r-ésima vez. La variable aleatoria que proporciona la probabilidad de que se produzcan k fracasos antes de obtener el r-ésimo éxito sigue una distribución binomial negativa de parámetros r y p, BN(r,p). La distribución geométrica corresponde al caso particular en que r=1. Un ejemplo es el número de lanzamientos fallidos de un dado antes de obtener un 6 en tres ocasiones, que sigue una BN(3,1/6).

En el caso de que los sucesos ocurran a intervalos regulares de tiempo, esta variable proporciona el tiempo total para que ocurran r éxitos, por lo que también se denomina "distribución binomial de tiempo de espera".

La distribución binomial negativa fue propuesta, originalmente, como una alternativa a la distribución de Poisson para modelar el número de ocurrencias de un suceso cuando los datos presentan lo que se conoce como variación extra-Poisson o sobredispersión. En estas situaciones, la varianza es mayor que la media, por lo que se incumple la propiedad que caracteriza a una distribución de Poisson, según la cual la media es igual a la varianza. La primera aplicación en bioestadística la realizó Student (William S. Gosset) a principios de siglo cuando propuso esta distribución para modelar el número de glóbulos rojos en una gota de sangre. En este caso, la variabilidad extra se debe al hecho de que esas células no están uniformemente distribuida en la gota, es decir, la tasa de intensidad no es homogénea.

Por ejemplo, la distribución binomial negativa es más adecuada que la de Poisson para

modelar el número de accidentes laborales ocurridos en un determinado lapso. La distribución de Poisson asume que todos los individuos tienen la misma probabilidad de sufrir un accidente y que ésta permanece constante durante el período de estudio; sin

embargo, es más plausible la hipótesis de que los individuos tienen probabilidades constantes en el tiempo, pero que varían de unos sujetos a otros; esto es lo que se conoce en la literatura como la propensión a los accidentes ("*accident proneness*")[8,9]. Esta hipótesis se traduce en una distribución de Poisson mixta, o de efectos aleatorios, en la que se supone que las probabilidades varían entre individuos de acuerdo a una distribución gamma y esto resulta en una distribución binomial negativa para el número de accidentes.

Valores:

x: 0, 1, 2, ...

Parámetros:

p: probabilidad de éxito, $0<p<1$

r: número de éxitos, $r\geq 0$

Ejercicio

Se sabe que, en promedio, de cada 100 placas de rayos X que se realizan, una es defectuosa.

¿Cuál es el número medio de placas útiles que se producen entre 10 defectuosas?

Si se considera el primer fallo como punto de inicio, hay que considerar la variable "número de placas útiles antes de 9 defectuosas", que sigue una distribución binomial negativa de parámetros r=9 y p=0,01.

Resultados con Epidat 3.1

```
Cálculo de probabilidades. Distribuciones discretas

Binomial negativa (r,p)
r : Número de éxitos                  9
p : Probabilidad de éxito        0,0100
Punto K                               1

Media                          891,0000
Varianza                      89100,0000
```

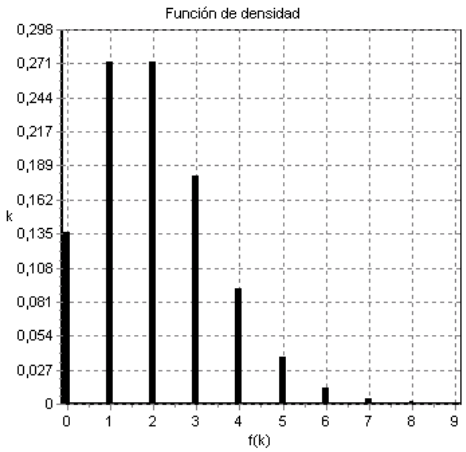
Entre 10 placas defectuosas se producen, en promedio, unas 891 placas útiles.

5.2.6. Distribución Poisson (lambda)

La distribución de Poisson, que debe su nombre al matemático francés Simeón Denis Poisson (1781-1840), ya había sido introducida en 1718 por Abraham De Moivre como una forma límite de la distribución binomial que surge cuando se observa un evento raro después de un número grande de repeticiones[10]. En general, la distribución de Poisson se puede utilizar como una aproximación de la binomial, Bin(n, p), si el número de pruebas n es grande, pero la probabilidad de éxito p es pequeña; una regla es que la aproximación Poisson-binomial es "buena" si $n\geq 20$ y $p\leq 0{,}05$ y "muy buena" si $n\geq 100$ y $p\leq 0{,}01$.

La distribución de Poisson también surge cuando un evento o suceso “raro” ocurre aleatoriamente en el espacio o el tiempo. La variable asociada es el número de ocurrencias del evento en un intervalo o espacio continuo, por tanto, es una variable aleatoria discreta que toma valores enteros de 0 en adelante (0, 1, 2,...). Así, el número de pacientes que llegan a un consultorio en un lapso dado, el número de llamadas que recibe un servicio de atención a urgencias durante 1 hora, el número de células anormales en una superficie histológica o el número de glóbulos blancos en un milímetro cúbico de sangre son ejemplos de variables que siguen una distribución de Poisson. En general, es una distribución muy utilizada en diversas áreas de la investigación médica y, en particular, en epidemiología.

El concepto de evento “raro” o poco frecuente debe ser entendido en el sentido de que la probabilidad de observar k eventos decrece rápidamente a medida que k aumenta. Supóngase, por ejemplo, que el número de reacciones adversas tras la administración de un fármaco sigue una distribución de Poisson de media *lambda*=2. Si se administra este fármaco a 1.000 individuos, la probabilidad de que se produzca una reacción adversa (k=1) es 0,27; los valores de dicha probabilidad para k=2, 3, 4, 5, 6 reacciones, respectivamente, son: 0,27; 0,18; 0,09; 0,03 y 0,01. Para k=10 o mayor, la probabilidad es virtualmente 0. El rápido descenso de la probabilidad de que se produzcan k reacciones adversas a medida que k aumenta puede observarse claramente en el gráfico de la función de densidad obtenido con Epidat



Para que una variable recuento siga una distribución de Poisson deben cumplirse varias

condiciones:

1. En un intervalo muy pequeño (p. e. de un milisegundo) la probabilidad de que

ocurra un evento es proporcional al tamaño del intervalo.

2. La probabilidad de que ocurran dos o más eventos en un intervalo muy pequeño es tan reducida que, a efectos prácticos, se puede considerar nula.

3. El número de ocurrencias en un intervalo pequeño no depende de lo que ocurra en

cualquier otro intervalo pequeño que no se solape con aquél.

Estas propiedades pueden resumirse en que el proceso que genera una distribución de

Poisson es estable (produce, a largo plazo, un número medio de sucesos constante por unidad de observación) y no tiene memoria (conocer el número de sucesos en un intervalo no ayuda a predecir el número de sucesos en el siguiente).

El parámetro de la distribución, *lambda*, representa el número promedio de eventos

esperados por unidad de tiempo o de espacio, por lo que también se suele hablar de *lambda* como “la tasa de ocurrencia” del fenómeno que se observa.

A veces se usan variables de Poisson con "intervalos" que no son espaciales ni temporales, sino de otro tipo. Por ejemplo, para medir la frecuencia de una enfermedad se puede contar, en un período dado, el número de enfermos en cierta población, dividida en "intervalos" de, por ejemplo, 10.000 habitantes. Al número de personas enfermas en una población de tamaño prefijado, en un instante dado, se le denomina prevalencia de la enfermedad en ese instante y es una variable que sigue una distribución de Poisson. Otra medida para la frecuencia de una enfermedad es la incidencia, que es el número de personas que enferman en una población en un periodo determinado. En este caso, el intervalo es de personas-tiempo, habitualmente personas-año, y es también una variable con distribución de Poisson. Habitualmente, ambas medidas se expresan para intervalos de tamaño unidad o, dicho de otro modo, en lugar de la variable número de enfermos, se usa el parámetro *lambda* (el riesgo, en el caso de la prevalencia, y la densidad de incidencia, en el de incidencia).

La distribución de Poisson tiene iguales la media y la varianza. Si la variación de los casos observados en una población excede a la variación esperada por la Poisson, se está ante la presencia de un problema conocido como sobredispersión y, en tal caso, la distribución binomial negativa es más adecuada.

Valores:

x: 0, 1, 2, ...

Parámetros:

lambda: media de la distribución, *lambda* > 0

Ejercicio

El número de enfermos que solicitan atención de urgencia en un hospital durante un periodo

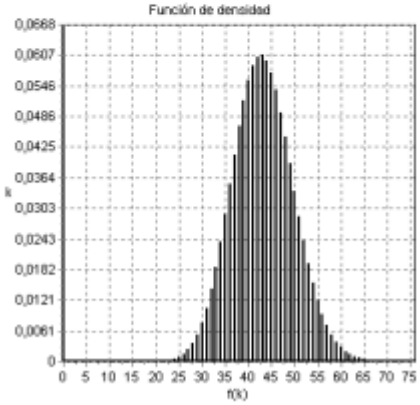
de 24 horas tiene una media de 43,2 pacientes. Se sabe que el servicio se colapsará si el número de enfermos excede de 50. ¿Cuál es la probabilidad de que se colapse el servicio de urgencias del hospital? Representar la función de densidad de probabilidad.

Para calcular la probabilidad pedida y, además, representar la función de densidad de

probabilidad hay que marcar el cuadro situado en la parte inferior derecha de la pantalla: *Obtener las funciones de distribución y densidad.*

Resultados con Epidat 3.1

```
Cálculo de probabilidades. Distribuciones discretas

Poisson (lambda)
lambda : Media                    43,2000
Punto K                                50

Probabilidad    Pr[X=k]             0,0339
Cola Izquierda  Pr[X<=k]            0,8657
Cola Derecha    Pr[X>k]             0,1343

Media                             43,2000
Varianza                          43,2000
```



La probabilidad de que el servicio colapse está cerca de 0,13.

La probabilidad de que el servicio colapse está cerca de 0,13.

5.2.7. Resumen

En modo de resumen les mostrare las formulas utilizadas para cada uno de los modelos de distribución

Algunas Distribuciones Discretas

$$MEDIA = \mu_x = \sum_x xf(x)$$

$$VARIANZA = E[(x-\mu)^2] = \sum_x (x-\mu)^2 \cdot f(x) = E[x^2] - \mu^2$$

Distribución Discreta Uniforme

$$f(x;k) = \frac{1}{k} \qquad x = x_1, x_2, \dots, x_k$$

$$\mu = \frac{\sum_{i=1}^{k} X_i}{k} \sigma^2 = \frac{\sum_{i=1}^{k}(X_i-\mu)^2}{k}$$

Distribución Binomial

$$b(x;n,p) = \binom{n}{x} p^x q^{n-x} \qquad x = 0,1,2,\dots,n$$

μ=np σ^2=npq

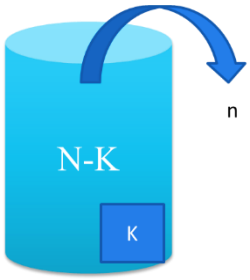
Distribución Multinomial

$$f(x_1, x_2, \dots, x_k;\ p_1, p_2, \dots, p_k) = \binom{n}{x_1, x_2, \dots, x_k} {p_1}^{x_1} \cdot {p_2}^{x_2} \cdot \dots \cdot {p_k}^{x_k}$$

$$\sum_{i=1}^{k} x_i = n \sum_{i=1}^{k} p_i = 1$$

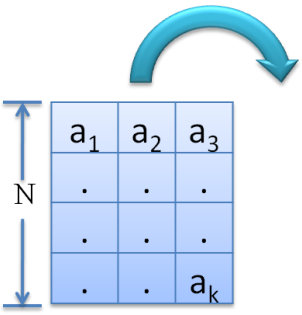
Distribución Hipergeométrica

$$h(x; N; n, k) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}, \qquad x = 0,1,2, \dots, n$$



$$\mu = n\frac{k}{N}\sigma^2 = \left(\frac{N-n}{N-1}\right) \cdot n \cdot \frac{k}{N} \cdot \left(1 - \frac{k}{N}\right)$$

Distribución Multivariada

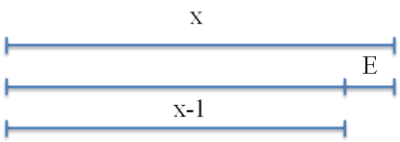


$$f(x_1, x_2, \dots, x_k;\ a_1, a_2, \dots, a_k) = \frac{\binom{a_1}{x_1}\binom{a_2}{x_2}\cdots\binom{a_k}{x_k}}{\binom{N}{n}}$$

$$\sum_{i=1}^{k} a_i = N \sum_{i=1}^{k} x_i = n$$

Distribución Binomial Negativa

$$b*(x; k, p) = \binom{x-1}{k-1} p^k q^{x-k}, \qquad x = k, k+1, k+2, ..$$



Distribución Geométrica

$$g(x; p) = p \cdot q^{x-1} \qquad x = 1,2,3 \dots$$

$$\mu = \frac{1}{p}\sigma^2 = \frac{1-p}{p^2}$$

Distribución De Poisson

$$p(x; \lambda t) = \frac{e^{-\lambda t}(\lambda t)^x}{x!} \qquad x = 01,2,3, \dots$$

$$\mu = -\lambda t$$

$$\sigma^2 = \lambda t$$

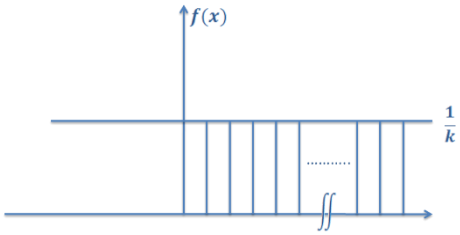
Distribución Uniforme Discreta

$$x = \{k\}$$

$$x = \{x_1, x_2, \dots, x_k\}$$

$$P[x_1] = \frac{1}{k} = \cdots = P[x_k] = \frac{1}{k}$$

$$f(x,k) = \frac{1}{k}$$



$$\mu = \mu_x = E[x] = \sum x_i \cdot \frac{1}{k} = \sum_{k=1} x_i$$

$$\sigma^2 = E[(x-\mu)^2] = \sum_{i=1}^{k} (x_i - \mu)^2 \cdot \frac{1}{k} = \frac{\sum_{i=1}^{k} (x_i - \mu)^2}{k}$$

6. ALGUNAS DISTRIBUCIONES CONTINUAS DE PROBABILIDAD

Las distribuciones continuas incluidas en el módulo de "Cálculo de probabilidades" son:

- Uniforme
- Normal
- Lognormal
- Logística
- Beta
- Gamma
- Exponencial
- Ji-cuadrado
- t de Student
- F de Snedecor

6.1.Distribución Uniforme (a,b)

La distribución uniforme es útil para describir una variable aleatoria con probabilidad constante sobre el intervalo [a,b] en el que está definida. Esta distribución presenta una peculiaridad importante: la probabilidad de un suceso dependerá exclusivamente de la amplitud del intervalo considerado y no de su posición en el campo de variación de la variable.

Cualquiera sea la distribución F de cierta variable X, la variable transformada Y=F(X) sigue una distribución uniforme en el intervalo [0,1]. Esta propiedad es fundamental por ser la base para la generación de números aleatorios de cualquier distribución en las técnicas de simulación

Campo de variación:

$a \leq x \leq b$ Parámetros:

a: mínimo del recorrido b: máximo del recorrido

Ejercicio

Supóngase una variable que se distribuye uniformemente entre 380 y 1.200. Determínese:

1. La probabilidad de que el valor de la variable sea superior a mil.

2. La media y la desviación estándar de dicha variable.

A Epidat se le proporcionará el límite superior e inferior del campo de variación de la variable [380, 1.200] y, además, el punto a partir del cual se quiere calcular la probabilidad.

Resultados con Epidat 3.1

```
Cálculo de probabilidades. Distribuciones continuas

Uniforme (a,b)
a : Mínimo                    380,0000
b : Máximo                      1200,0000
Punto X                         1000,0000

Cola Izquierda Pr[X<=k]            0,7561
Cola Derecha   Pr[X>=k]            0,2439

Media                            790,0000
Varianza                        56033,3333
Mediana                          790,0000
```

La probabilidad de que la variable sea superior a mil se sitúa en un entorno de 0,24, la media es 790 y la desviación estándar, raíz cuadrada de la varianza, es aproximadamente 237.

Ejercicio

Un contratista A está preparando una oferta sobre un nuevo proyecto de construcción. La oferta sigue una distribución uniforme entre 55 y 75 miles de euros. Determínese:

1. La probabilidad de que la oferta sea superior a 60 mil euros.
2. La media y la desviación estándar de la oferta.

A Epidat se le proporcionará el límite superior e inferior del campo de variación de la variable [55, 75] y, además, el punto a partir del cual se quiere calcular la probabilidad.

Resultados con Epidat 3.1

```
Cálculo de probabilidades. Distribuciones continuas

Uniforme (a,b)
a : Mínimo                      55,0000
b : Máximo                      75,0000
Punto X                         60,0000

Cola Izquierda   Pr[X<=k]        0,2500
Cola Derecha     Pr[X>=k]        0,7500


Media                           65,0000
Varianza                        33,3333
Mediana                         65,0000
```

La probabilidad de que la oferta sea superior a 60 mil euros se sitúa en un entorno de 0,75, y la media es 65.

6.2.Distribución Normal (Mu, Sigma)

La distribución normal es, sin duda, la distribución de probabilidad más importante del Cálculo de probabilidades y de la Estadística. Fue descubierta por De Moivre (1773), como aproximación de la distribución binomial. De todas formas, la importancia de la distribución

normal queda totalmente consolidada por ser la distribución límite de numerosas variables aleatorias, discretas y continuas, como se demuestra a través de los teoremas centrales del límite. Las consecuencias de estos teoremas implican la casi universal presencia de la distribución normal en todos los campos de las ciencias empíricas: biología, medicina, psicología, física, economía, etc. En particular, muchas medidas de datos continuos en medicina y en biología (talla, presión arterial, etc.) se aproximan a la distribución normal.

Junto a lo anterior, no es menos importante el interés que supone la simplicidad de sus características y de que de ella derivan, entre otras, tres distribuciones (Ji-cuadrado, t y F) que se mencionarán más adelante, de importancia clave en el campo de la contrastación de hipótesis estadísticas.

La distribución normal queda totalmente definida mediante dos parámetros: la media (*Mu*) y la desviación estándar (*Sigma*).

Campo de variación:

$-\infty < x < \infty$

Parámetros:

Mu: media de la distribución, $-\infty < Mu < \infty$

Sigma: desviación estándar de la distribución, $Sigma > 0$

Ejercicio

Se supone que el nivel de colesterol de los enfermos de un hospital sigue una distribución normal con una media de 179,1 mg/dL y una desviación estándar de 28,2 mg/dL.

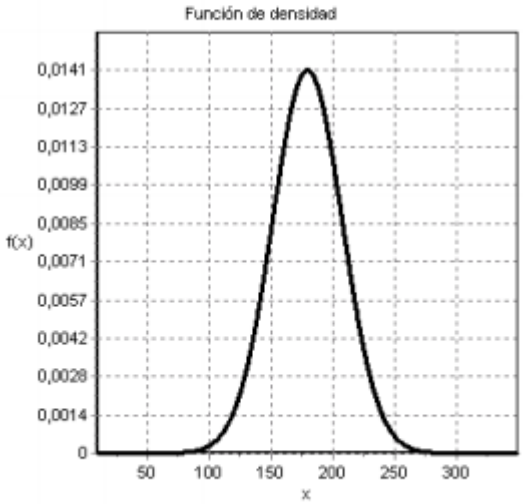
1. Calcule el porcentaje de enfermos con un nivel de colesterol inferior a 169 mg/dL.

2. ¿Cuál será el valor del nivel de colesterol a partir del cual se encuentra el 10% de los enfermos del hospital con los niveles más altos?

3. Represente la función de densidad.

En este caso, se tendrá que ejecutar Epidat 3.1 dos veces: en el primer caso para calcular una probabilidad, en el segundo caso el dato de entrada es una probabilidad, concretamente la cola de la derecha, lo que permitirá obtener el punto. En ambas ejecuciones se ofrece, de manera opcional, la función de densidad del nivel de colesterol.

```
Cálculo de probabilidades. Distribuciones continuas

Normal (Mu,Sigma)
Mu    : Media                    179,1000
Sigma : Desviación estándar       28,2000
Punto X                          169,0000

Cola Izquierda Pr[X<=k]            0,3601
Cola Derecha   Pr[X>=k]            0,6399
Dos Colas      1-Pr[|X|<=k]        0,7202
```



El porcentaje de enfermos con un nivel de colesterol inferior a 169 mg/dL es 36%.

2. Resultados con Epidat 3.1

```
Cálculo de probabilidades. Distribuciones continuas
Normal (Mu,Sigma)
Mu    : Media                    179,1000
Sigma : Desviación estándar       28,2000
Cola Izquierda Pr[X<=k]            0,9000
Cola Derecha   Pr[X>=k]            0,1000
Dos Colas      1-Pr[|X|<=k]        0,2000

Punto X                          215,2398
```

A partir de 215,24 mg/dL se encuentran los valores de colesterol del 10% de los enfermos que tienen los valores más altos.

6.3.Distribución Lognormal (Mu, Sigma)

La variable resultante al aplicar la función exponencial a una variable que se distribuye

normal con media Mu y desviación estándar Sigma, sigue una distribución lognormal con parámetros Mu (escala) y Sigma (forma). Dicho de otro modo, si una variable X se distribuye normalmente, la variable lnX, sigue una distribución lognormal.

La distribución lognormal es útil para modelar datos de numerosos estudios médicos tales

como el período de incubación de una enfermedad, los títulos de anticuerpo a un virus, el tiempo de supervivencia en pacientes con cáncer o SIDA, el tiempo hasta la seroconversión de VIH+, etc.

Campo de variación:

$0 < x < \infty$ Parámetros:

Mu: parámetro de escala, $-\infty < Mu < \infty$ Sigma: parámetro de forma, Sigma > 0

Ejercicio Supóngase que la supervivencia, en años, luego de una intervención quirúrgica (tiempo que pasa hasta que ocurre la muerte del enfermo) en una cierta población sigue una distribución lognormal de parámetro de escala 2,32 y de forma 0,20. Calcúlese la probabilidad de supervivencia a los 12 años, la mediana de supervivencia y represente la función de distribución de la variable.

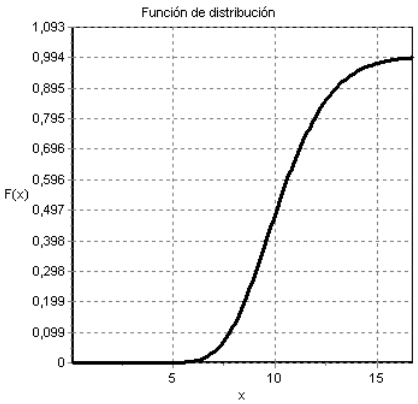
Resultados con Epidat

Resultados con Epidat 3.1

```
Cálculo de probabilidades. Distribuciones continuas

Lognormal (Mu,Sigma)
Mu    : Escala                  2,3200
Sigma : Forma                   0,2000
Punto X                        12,0000

Cola Izquierda Pr[X<=k]         0,7952
Cola Derecha   Pr[X>=k]         0,2048

Media                          10,3812
Varianza                        4,3982
Mediana                        10,1757
Moda                            9,7767
```

La probabilidad de supervivencia a los 12 años se sitúa próximo a 0,20.

La función de distribución de la supervivencia a la intervención quirúrgica se presenta a continuación:



6.4.Distribución Logística (a, b)

La distribución logística se utiliza en el estudio del crecimiento temporal de variables, en particular, demográficas. En biología se ha aplicado, por ejemplo, para modelar el crecimiento de células de levadura, y para representar curvas de dosis-respuesta en bioensayos.

La más conocida y generalizada aplicación de la distribución logística en Ciencias de la Salud se fundamenta en la siguiente propiedad: si U es una variable uniformemente distribuida

en el intervalo [0,1], entonces la variable $X = Ln\left(\frac{U}{1-U}\right)$ sigue una distribución logística. Esta transformación, denominada logit, se utiliza para modelar datos de respuesta binaria, especialmente en el contexto de la regresión logística.

Campo de variación:

$-\infty < x < \infty$ Parámetros:

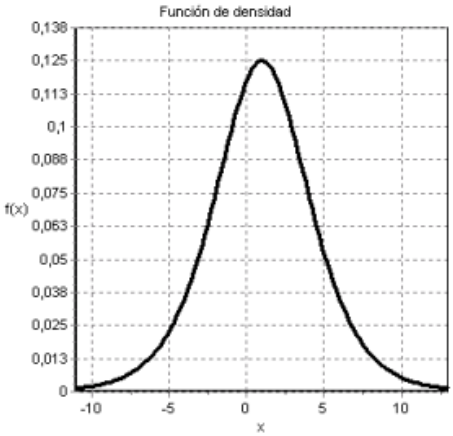
a: parámetro de posición, $-\infty < a < \infty$ b: parámetro de escala, $b > 0$

Ejercicio

El crecimiento relativo anual (%) de la población de un determinado país sigue una distribución logística de parámetro de posición 1 y de escala 2. Calcular la probabilidad de que el crecimiento en un año determinado sea superior al 5% y representar la función de densidad.

Resultados con Epidat 3.1

```
Cálculo de probabilidades. Distribuciones continuas
Logística (a,b)
a : Posición                    1,0000
b : Escala                      2,0000
Punto X                         5,0000

Cola Izquierda Pr[X<=k]         0,8808
Cola Derecha   Pr[X>=k]         0,1192
Dos Colas      1-Pr[|X|<=k]     0,2384
Media                           1,0000
Varianza                       13,1595
Mediana                         1,0000
Moda                            1,0000
```



La probabilidad de que la población tenga un crecimiento superior al 5% es del orden de 0,12.

6.5.Distribución Beta (p,q)

La distribución beta es posible para una variable aleatoria continua que toma valores en el intervalo [0,1], lo que la hace muy apropiada para modelar proporciones. En la inferencia bayesiana, por ejemplo, es muy utilizada como distribución a priori cuando las observaciones tienen una distribución binomial.

Uno de los principales recursos de esta distribución es el ajuste a una gran variedad de distribuciones empíricas, pues adopta formas muy diversas dependiendo de cuáles sean los valores de los parámetros de forma p y q, mediante los que viene definida la distribución.

Un caso particular de la distribución beta es la distribución uniforme en [0,1], que se corresponde con una beta de parámetros p=1 y q=1, denotada Beta(1,1).

Campo de variación:

$0 \leq x \leq 1$ Parámetros:

p: parámetro de forma, $p > 0$ q: parámetro de forma, $q > 0$

Ejercicio

En el presupuesto familiar, la porción que se dedica a salud sigue una distribución Beta(2,2).

1. ¿Cuál es la probabilidad de que se gaste más del 25% del presupuesto familiar en salud?

2. ¿Cuál será el porcentaje medio que las familias dedican a la compra de productos y servicios de salud?

Resultados con Epidat 3.1

```
Cálculo de probabilidades. Distribuciones continuas

Beta (p,q)
p : Forma                        2,0000
q : Forma                        2,0000
Punto X                          0,2500

Cola Izquierda Pr[X<=k]          0,1563
Cola Derecha   Pr[X>=k]          0,8438

Media                            0,5000
Varianza                         0,0500
Moda                             0,5000
```

Teniendo en cuenta la distribución beta, la probabilidad de que se gaste más de la cuarta parte del presupuesto en salud será 0,84 y el porcentaje medio que las familias dedican a la compra de productos y servicios de salud será el 50%.

6.6.Distribución Gamma (a,p)

La distribución gamma se puede caracterizar del modo siguiente: si se está interesado en la ocurrencia de un evento generado por un proceso de Poisson de media lambda, la variable que mide el tiempo transcurrido hasta obtener n ocurrencias del evento sigue una distribución gamma con parámetros a= n×lambda (escala) y p=n (forma). Se denota Gamma(a,p).

Por ejemplo, la distribución gamma aparece cuando se realiza el estudio de la duración de elementos físicos (tiempo de vida).

Esta distribución presenta como propiedad interesante la “falta de memoria”. Por esta razón,

es muy utilizada en las teorías de la fiabilidad, mantenimiento y fenómenos de espera (por ejemplo en una consulta médica “tiempo que transcurre hasta la llegada del segundo paciente”).

Campo de variación:

$0 < x < \infty$ Parámetros:

a: parámetro de escala, $a > 0$ p: parámetro de forma, $p > 0$

Ejercicio 1

El número de pacientes que llegan a la consulta de un médico sigue una distribución de

Poisson de media 3 pacientes por hora. Calcular la probabilidad de que transcurra menos de una hora hasta la llegada del segundo paciente.

Debe tenerse en cuenta que la variable aleatoria “tiempo que transcurre hasta la llegada del

segundo paciente” sigue una distribución Gamma (6, 2).

```
Cálculo de probabilidades. Distribuciones continuas

Gamma (a,p)
a : Escala                    6,0000
p : Forma                     2,0000
Punto X                       1,0000

Cola Izquierda Pr[X<=k]       0,9826
Cola Derecha   Pr[X>=k]       0,0174

Media                         0,3333
Varianza                      0,0556
Moda                          0,1667
```

La probabilidad de que transcurra menos de una hora hasta que llegue el segundo paciente

es 0,98.

Ejercicio 2

Suponiendo que el tiempo de supervivencia, en años, de pacientes que son sometidos a una

cierta intervención quirúrgica en un hospital sigue una distribución Gamma con parámetros a=0,81 y p=7,81, calcúlese:

1. El tiempo medio de supervivencia.

2. Los años a partir de los cuales la probabilidad de supervivencia es menor que 0,1.

Resultados con Epidat

Resultados con Epidat 3.1

```
Cálculo de probabilidades. Distribuciones continuas

Gamma (a,p)
a : Escala                    0,8100
p : Forma                     7,8100
Cola Izquierda Pr[X<=k]       0,9000
Cola Derecha   Pr[X>=k]       0,1000

Punto X                       14,2429

Media                         9,6420
Varianza                      11,9037
Moda                          8,4074
```

El tiempo medio de supervivencia es de, aproximadamente, 10 años.

6.7.Distribución Exponencial (lambda)

La distribución exponencial es el equivalente continuo de la distribución geométrica discreta.

Esta ley de distribución describe procesos en los que interesa saber el tiempo hasta que ocurre determinado evento; en particular, se utiliza para modelar tiempos de supervivencia. Un ejemplo es el tiempo que tarda una partícula radiactiva en desintegrarse. El conocimiento de la ley que sigue este evento se utiliza, por ejemplo, para la datación de fósiles o cualquier materia orgánica mediante la técnica del carbono 14.

Una característica importante de esta distribución es la propiedad conocida como “falta de

memoria”. Esto significa, por ejemplo, que la probabilidad de que un individuo de edad t sobreviva x años más, hasta la edad x+t, es la misma que tiene un recién nacido de sobrevivir hasta la edad x. Dicho de manera más general, el tiempo transcurrido desde cualquier instante dado t0 hasta que ocurre el evento, no depende de lo que haya ocurrido antes del instante t0.

La distribución exponencial se puede caracterizar como la distribución del tiempo entre

sucesos consecutivos generados por un proceso de Poisson; por ejemplo, el tiempo que transcurre entre dos heridas graves sufridas por una persona. La media de la distribución de Poisson, lambda, que representa la tasa de ocurrencia del evento por unidad de tiempo, es el parámetro de la distribución exponencial, y su inversa es el valor medio de la distribución.

También se puede ver como un caso particular de la distribución gamma(a,p), con a=lambda y p=1. El uso de la distribución exponencial ha sido limitado en bioestadística, debido a la propiedad de falta de memoria que la hace demasiado restrictiva para la mayoría de los problemas.

Campo de variación:

$0 < x < \infty$ Parámetros:

lambda: tasa, lambda > 0

Ejercicio

Se ha comprobado que el tiempo de vida de cierto tipo de marcapasos sigue una distribución

exponencial con media de 16 años. ¿Cuál es la probabilidad de que a una persona a la que se le ha implantado este marcapasos se le deba reimplantar otro antes de 20 años? Si el marcapasos lleva funcionando correctamente 5 años en un paciente, ¿cuál es la probabilidad de que haya que cambiarlo antes de 25 años?

La variable aleatoria “tiempo de vida del marcapasos” sigue una distribución exponencial de parámetro lambda=1/16=0,0625

Resultados con Epidat 3.1

```
Cálculo de probabilidades. Distribuciones continuas

Exponencial (lambda)
lambda : Tasa                    0,0625
Punto X                         20,0000

Cola Izquierda Pr[X<=k]          0,7135
Cola Derecha   Pr[X>=k]          0,2865
```

La probabilidad de que se le tenga que implantar otro marcapasos antes de los 20 años se sitúa en un entorno a 0,71.

Teniendo en cuenta la propiedad de “falta de memoria” de la exponencial, la probabilidad

de tener que cambiar antes de 25 años un marcapasos que lleva funcionando 5 es igual a la probabilidad de cambio a los 20 años, es decir, P(X<25/X>5) = P(X<20) = 0,71.

6.8.Distribución Ji-cuadrado (n)

Un caso especial, muy importante, de la distribución Gamma se obtiene cuando a=1/2 y

p=n/2. La distribución resultante se conoce con el nombre de Ji-cuadrado con n grados de libertad. Es la distribución que sigue la suma de los cuadrados de n variables independientes N(0,1).

La Ji-cuadrado es una distribución fundamental en inferencia estadística y en los tests

estadísticos de bondad de ajuste. Se emplea, entre muchas otras aplicaciones, para determinar los límites de confianza de la varianza de una población normal, para contrastar la hipótesis de homogeneidad o de independencia en una tabla de contingencia y para pruebas de bondad de ajuste.

La distribución Ji-cuadrado queda totalmente definida mediante sus grados de libertad n. Campo de variación:

$0 \leq x < \infty$ Parámetros:

n: grados de libertad, n>0

Ejercicio

Considere la distribución Ji-cuadrado con 2 grados de libertad.

1. ¿Qué proporción del área bajo la curva se ubica a la derecha de 9,21?

2. ¿Qué valor de la variable aísla el 10% superior de la distribución?

1. Resultados con Epidat 3.1

```
Cálculo de probabilidades. Distribuciones continuas

Ji-cuadrado (n)
n : Grados de libertad                2
Punto X                          9,2100

Cola Izquierda Pr[X<=k]          0,9900
Cola Derecha   Pr[X>=k]          0,0100
```

El 1% del área bajo la curva se ubica a la derecha de 9,21.

2. Resultados con Epidat 3.1

```
Cálculo de probabilidades. Distribuciones continuas

Ji-cuadrado (n)
n : Grados de libertad                2
Cola Izquierda Pr[X<=k]          0,9000
Cola Derecha   Pr[X>=k]          0,1000

Punto X                          4,6052
```

El valor 4,6052 divide a la distribución en dos partes: el 90% de ésta queda a la izquierda de dicho punto y el 10% a la derecha.

6.9.Distribución t de Student (n)

La distribución t de Student se construye como un cociente entre una normal y la raíz de una

Ji-cuadrado independientes. Esta distribución desempeña un papel importante en la inferencia estadística asociada a la teoría de muestras pequeñas. Se usa habitualmente en el contraste de hipótesis para la media de una población, o para comparar las medias de dos poblaciones, y viene definida por sus grados de libertad n.

A medida que aumentan los grados de libertad, la distribución t de Student se aproxima a

una normal de media 0 y varianza 1 (normal estándar).

Campo de variación:

$-\infty < x < \infty$ Parámetros:

n: grados de libertad, n>0

Ejercicio

La distribución t de Student se aproxima a la normal a medida que aumentan los grados de

libertad.

- ✓ Calcular, para una distribución N(0,1), el punto que deja a la derecha una cola de probabilidad 0,05.
- ✓ Calcular, para una distribución t de Student, la probabilidad de que la variable tome un valor a la derecha de ese punto. Tomar como grados de libertad sucesivamente n= 10 y n= 500.

Para el primer apartado hay que seleccionar en la lista de distribuciones la normal de

parámetros Mu=0 y Sigma=1.

1. Resultados con Epidat 3.1

1. Resultados con Epidat 3.1

```
Cálculo de probabilidades. Distribuciones continuas

Normal (Mu,Sigma)
Mu    : Media                    0,0000
Sigma : Desviación estándar      1,0000
Cola Izquierda Pr[X<=k]          0,9500
Cola Derecha   Pr[X>=k]          0,0500
Dos Colas      1-Pr[|X|<=k]      0,1000

Punto X                          1,6449

Media                            0,0000
Varianza                         1,0000
```

En el segundo apartado se ejecutará dos veces Epidat 3.1: la primera vez con una distribución t de Student con 10 grados de libertad y la segunda vez con 500 grados de libertad.

2. Resultados con Epidat

```
Cálculo de probabilidades. Distribuciones continuas

t de Student (n)
n : Grados de libertad              10
Punto X                          1,6449

Cola Izquierda Pr[X<=k]          0,9345
Cola Derecha   Pr[X>=k]          0,0655
Dos Colas      1-Pr[|X|<=k]      0,1310
```

```
Cálculo de probabilidades. Distribuciones continuas

t de Student (n)
n : Grados de libertad             500
Punto X                          1,6449

Cola Izquierda Pr[X<=k]          0,9497
Cola Derecha   Pr[X>=k]          0,0503
Dos Colas      1-Pr[|X|<=k]      0,1006
```

Se aprecia claramente que, al aumentar los grados de libertad de la t de Student, la probabilidad se acerca a la calculada con la distribución Normal.

6.10. Distribución F de Snedecor (n,m)

Otra de las distribuciones importantes asociadas a la normal es la que se define como el cociente de dos variables con distribución Ji-cuadrado divididas por sus respectivos grados de libertad, n y m. En este caso la variable aleatoria sigue una distribución F de Snedecor de parámetros n y m. Hay muchas aplicaciones de la F en estadística y, en particular, tiene un papel importante en las técnicas del análisis de la varianza y del diseño de experimentos.

Campo de variación:

$0 \leq x < \infty$ Parámetros:

n: grados de libertad del numerador, n>0

m: grados de libertad del denominador, m>0

Ejercicio

En un laboratorio se efectuaron ciertas mediciones y se comprobó que seguían una distribución F con 10 grados de libertad en el numerador y 12 grados de libertad en el denominador.

1. Calcule el valor que deja a la derecha el 5% del área bajo la curva de densidad.
2. ¿Cuál es la probabilidad de que la medición sea superior a 4,30?
3. Represente la función de distribución y de densidad de las medidas.

1. Resultados con Epidat

```
Cálculo de probabilidades. Distribuciones continuas

F de Snedecor (n,m)
n : Grados libertad del num.   10,0000
m : Grados libertad del denom.  12,0000
Cola Izquierda Pr[X<=k]         0,9500
Cola Derecha   Pr[X>=k]         0,0500

Punto X                         2,7534
```

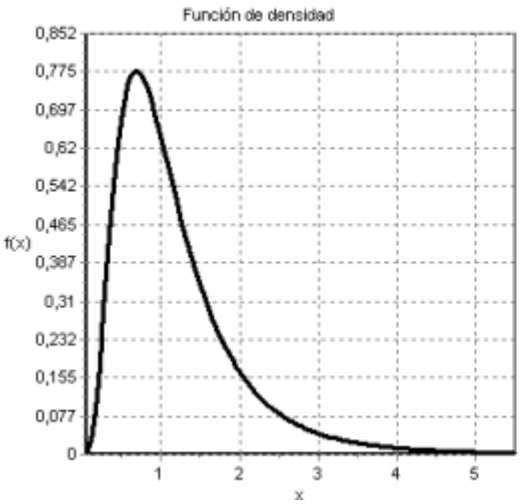
El valor que deja a la derecha una probabilidad de 0,05 es 2,75.

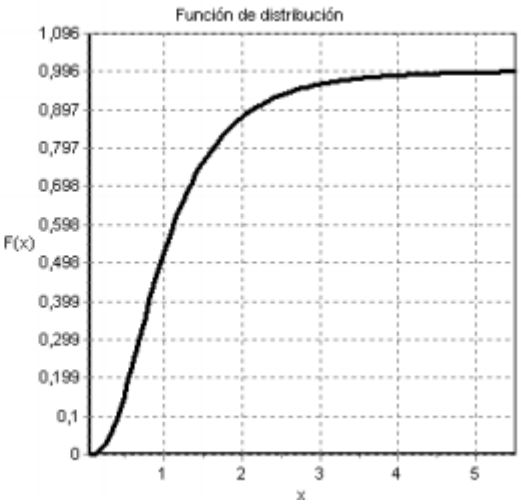
2. Resultados con Epidat

```
Cálculo de probabilidades. Distribuciones continuas

F de Snedecor (n,m)
n : Grados libertad del num.   10,0000
m : Grados libertad del denom.  12,0000
Punto X                         4,3000

Cola Izquierda Pr[X<=k]         0,9900
Cola Derecha   Pr[X>=k]         0,0100

Media                           1,2000
Varianza                        0,7200
Moda                            0,6857
```

La probabilidad que deja a la derecha 4,30 es 0,01.

3. Las funciones de densidad y distribución de las medidas efectuadas se presentan a continuación:





7. FUNCIONES DE VARIABLES ALEATORIAS

7.1. Funciones de variables aleatorias cambio de variable

7.1.1. Teorema

Si $X : (\Omega, A, P) \longrightarrow (R, B, P_X)$ es una variable aleatoria y $g : (R, B) \longrightarrow (R, B)$ es una función medible, entonces $Y = g(X) : (\Omega, A, P) \longrightarrow (R, B)$ es una variable aleatoria.

Demostración

$$\forall B \in B, \quad Y \quad (B) = (g(X))^{-1}(B) = X^{-1}(g^{-1}(B)) \in A \in B$$

Por tanto, Y es medible y, puesto que está definida sobre un espacio de probabilidad es una variable aleatoria.

Nos planteamos el problema de obtener la distribución de probabilidad de Y a partir de la de X. En teoría, el problema se resuelve de forma inmediata mediante el siguiente teorema general.

7.1.2. Teorema general de Cambio de Variable

Sea $X : (\Omega, A, P) \longrightarrow (R, B, P_X)$ una variable aleatoria y $g : (R, B) \longrightarrow (R, B)$ una función medible. Sea $Y = g(X)$, entonces

$$\forall B \in B, \qquad (B) = P_X(g^{-1}(B))$$

$$\forall y \in R, \quad F_Y(y) = P_X(g^{-1}((-\infty, y]))$$

<u>Demostración</u>

$$P_Y(B) = P[Y \quad (B)] = P[X^{-1}(g^{-1}(B))] = P_X(g^{-1}(B))$$

$$F_Y(y) = P_Y((-\infty, y]) = P_X(g^{-1}((-\infty, y]))$$

Pero en la práctica trabajaremos con variables discretas o continuas, o sea, con funciones masa de probabilidad o funciones de densidad. Nos interesa, por tanto, especificar las fórmulas de cambio de variable para tales casos.

7.2. Cambio de variable discreto

7.2.1. Teorema

Si $X : (\Omega, A, P) \longrightarrow (R, B, P_X)$ una variable aleatoria discreta, $P(\{X \in E_X\}) = 1$ y $g : (R, B) \longrightarrow (R, B)$ una función medible, entonces $Y = g(X)$ es discreta con valores en $g(E_X)$ y

$$P\{Y = y\} = \sum_{x \in g^{-1}(y) \cap E_X} P\{X = x\}, \quad \forall y \in g(E_X)$$

Demostración

$g(E_X)$ es numerable y $P\{Y \in g(E_X)\} = P\{g(X) \in g(E_X)\} \geq P\{X \in E_X\} = 1$

$$\forall y \in g(E_X),\ P\{Y = y\} = P\{g(X) = y\} = P\{X \in g^{-1}(y)\} = \underset{x \in g^{-1}(y) \cap E_X}{P\{X = x\}}$$

7.3. Cambio De Variable Continuo

Ejemplo 1

Sea X una variable aleatoria continua con función de densidad

$$f(x) = \begin{cases} 1/2 & x \in (-1, 1) \\ 0 & x \notin (-1, 1) \end{cases}$$

y $g : (R, B) \longrightarrow (R, B)$ una función medible definida como

$$g(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases}$$

Ejemplo 2

$$g(x) = \begin{cases} x & x < 0 \\ 1 & x \geq 0 \end{cases}$$

Aquí analizaremos el caso de transformaciones que convierten una variable continua en discreta y el caso de transformaciones biunívocas de variables continuas (que transforman variables continuas en continuas).

7.3.1. Teorema de cambio de variable de continuo a discreto

Sea $X : (\Omega, A, P) \longrightarrow (R, B, P_X)$ una variable aleatoria continua con función de densidad f_X y sea $Y = g(X)$ una variable aleatoria discreta con valores en un conjunto numerable E_Y. Entonces,

$$\forall y \in E_Y, \quad P\{Y = y\} = \int_{g^{-1}(y)} f_X(x)\, dx$$

Demostración

$$P\{Y = y\} = P\{g(X) = y\} = P\{X \in g^{-1}(y)\} = \int_{g^{-1}(y)} f_X(x)\, dx$$

Ejemplo

Sea X una variable aleatoria continua con función de densidad

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad x \in \mathbb{R}$$

y sea Y una variable aleatoria definida como

$$Y = \begin{cases} 1 & X > 0 \\ 1/2 & X = 0 \\ -1 & X < 0 \end{cases}$$

7.3.2. Teorema de cambio de variable continuo a continuo

Sea $X : (\Omega, \mathcal{A}, P) \longrightarrow (\mathrm{R}, \mathcal{B}, P_X)$ una variable aleatoria continua con valores en el intervalo (a, b) ($f_X > 0 \iff x \in (a, b)$) y sea $g : (a, b) \longrightarrow \mathrm{R}$ una función derivable y estríctamente monótona (creciente o decreciente). Entonces $Y = g(X)$ es una variable aleatoria continua con función de densidad

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right| & y \in g((a, b)) \\ 0 & y \notin g((a, b)) \end{cases}$$

EJEMPLOS

Ejemplo 1

Sea X una variable aleatoria continua con función de densidad

$$f(x) = \begin{cases} 0 & x < 0 \\ \lambda e^{-\lambda x} & x \geq 0 \end{cases}$$

Calcular la función de densidad de $Y = X^2$.

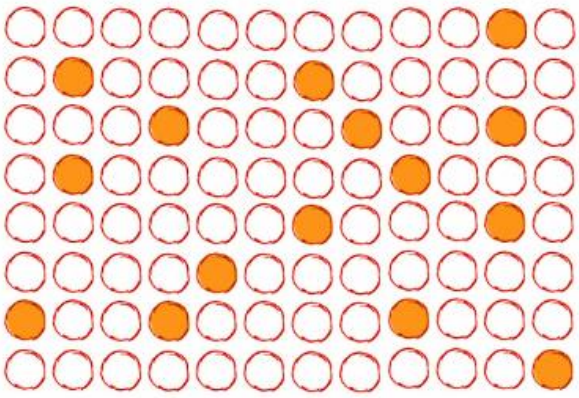
Ejemplo 2

Sea X una variable aleatoria continua con función de densidad f_X y función de distribución F_X. Calcular la función de densidad y la función de distribución de $Y = aX + b$, siendo $a, b \in \mathrm{R}$.

8. DISTRIBUCIONES FUNDAMENTALES DE MUESTREO Y DESCRIPCIONES DE DATOS

8.1. Muestreo aleatorio

El muestreo aleatorio constituye una de las clases más populares de muestreo aleatorio o probabilístico.



En esta técnica, cada miembro de la población tiene la misma probabilidad de ser seleccionado como sujeto. Todo el proceso de toma de muestras se realiza en un paso, en donde cada sujeto es seleccionado independientemente de los otros miembros de la población.

El muestreo aleatorio siemple se puede aplicar en muchos métodos. El más primitivo y mecánico sería el de la lotería. A cada miembro de la población se le asigna un número. Todos los números se colocan en un recipiente o un sombrero y se mezclan. Con los ojos vendados, el investigador va sacando las etiquetas con números. Todos los individuos que tengan los números sacados por el investigador son los sujetos del estudio. Otra forma sería que una computadora haga la selección al azar de la población. En el caso de poblaciones con pocos miembros, es aconsejable utilizar el primer método, pero si la población tiene muchos miembros, es preferible una selección aleatoria por computadora.

8.1.1. Ventajas del muestreo aleatorio simple

Una de las mejores cosas del muestreo aleatorio simple es la facilidad para armar la muestra. También se considera una forma justa de seleccionar una muestra a partir de una población, ya que cada miembro tiene igualdad de oportunidades de ser seleccionado.

Otra característica clave del muestreo aleatorio simple es la representatividad de la población. En teoría, lo único que puede poner en peligro su representatividad es la suerte. Si la muestra no es representativa de la población, la variación aleatoria es denominada error de muestreo.

Para sacar conclusiones de los resultados de un estudio son importantes una selección aleatoria imparcial y una muestra representatva. Recuerda que uno de los objetivos de la investigación es sacar conclusiones con relación a la población a partir de los resultados de una muestra. Debido a la representatividad de una muestra obtenida mediante un muestreo aleatorio simple, es razonable hacer generalizaciones a partir de los resultados de la muestra con respecto a la población.

8.1.2. Desventajas del muestreo aleatorio simple

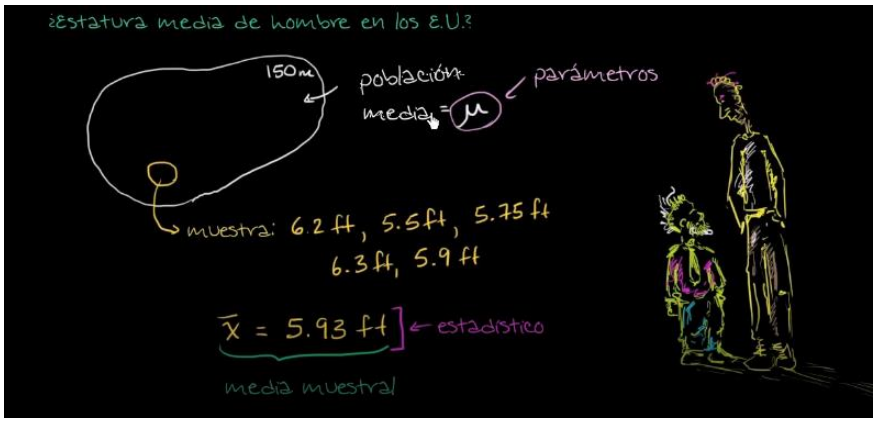
Una de las limitaciones más evidentes del muestreo aleatorio simple es la necesidad de una lista completa de todos los miembros de la población. Debes tener en cuenta que la lista de la población debe estar completa y actualizada. Esta lista generalmente no está disponible en poblaciones grandes. En estos casos, es más prudente utilizar otras técnicas de muestreo.

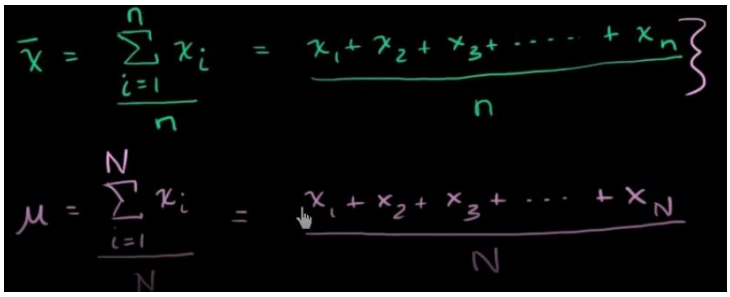
8.2. Inferencia de la media poblacional a partir de la media muestral

La Inferencia Estadística es aquella rama de la Estadística mediante la cual se trata de sacar conclusiones de una población en estudio, a partir de la información que proporciona una muestra representativa de la misma. También es denominada Estadística Inductiva o Inferencia Inductiva ya que es un procedimiento para generar nuevo conocimiento científico.

La muestra se obtiene por observación o experimentación. La necesidad de obtener un subconjunto reducido de la población es obvia si tenemos en cuenta los costes económicos de la experimentación o el hecho de que muchos de los métodos de medida son destructivos.

Toda inferencia inductiva exacta es imposible ya que disponemos de información parcial, sin embargo es posible realizar inferencias inseguras y medir el grado de inseguridad si el experimento se ha realizado de acuerdo con determinados principios. Uno de los propósitos de la inferencia Estadística es el de conseguir técnicas para hacer inferencias inductivas y medir el grado de incertidumbre de tales inferencias. La medida de la incertidumbre se realiza en términos de probabilidad.





La inferencia Estadística puede dividirse en dos apartados de acuerdo con el conocimiento sobre la distribución en la población.

8.2.1. Inferencia Paramétrica

Se conoce la forma de la distribución (Normal, Binomial, Poisson, etc ....) pero se desconocen sus parámetros. Se realizan inferencias sobre los parámetros desconocidos de la distribución conocida.

8.2.2. Inferencia No Parámetrica

Forma y parámetros desconocidos. Se realizan inferencias sobre características que no tienen porque ser parámetros de una distribución conocida (Mediana, Estadísticos de Orden).

De acuerdo con la forma en que se estudian los parámetros o características desconocidas, la inferencia puede dividirse en dos apartados:

8.2.3. Estimación

Se intenta dar estimaciones de los parámetros desconocidos sin hacer hipótesis previas sobre posibles valores de los mismos.

- ✓ **Estimación puntual**: Un único valor para cada parámetro.
- ✓ **Estimación por intervalos: Intervalo de valores probables para el parámetro.**
- ✓ **Contraste de Hipótesis:** Se realizan hipótesis sobre los parámetros desconocidos y se desarrolla un procedimiento para comprobar la verosimilitud de la hipótesis planteada.

En la mayor parte de las investigaciones reales suponemos que las variables o transformaciones de las mismas (logaritmos, etc, ...) tienen distribuciones aproximadamente normales.

8.2.4. Teorema del limite central

Lo que hemos visto hasta el momento parece bastante restrictivo ya que hemos supuesto, de entrada, que la distribución en la población es normal, pero existen muchos casos en los que no es posible suponer distribución Normal. El siguiente resultado permite trabajar con la normal para la distribución muestral de medias aunque la población no lo sea, y es conocido como Teorema Central del Límite.

Sea X1, X2, ... , Xn , una muestra aleatoria de una población X con una distribución de probabilidad no especificada para la que la media es E(X) = m y la varianza Var(X)= s2 finita. La media muestral tiene una distribución con media m y varianza s2 /n que tiende a una distribución normal cuando n tiende a infinito.

La demostración del resultado excede los límites de un curso introductorio.

La aproximación a la distribución normal es mejor para n grande ya que se trata de una aproximación y no de una distribución exacta como en el caso de poblaciones normales. En Estadística consideramos n grande cuando es mayor de 30.

Una consecuencia directa del teorema es que la suma de los valores muestrales sigue una distribución normal de media nm y varianza ns2.

El teorema de De Moivre que se explicó en el apartado de la normal puede entenderse también como un caso particular del Teorema Central del Límite.

Sea una población en la que se mide una v.a. X con distribución binomial B(1,p), es decir, toma el valor 1 con probabilidad p y el valor 0 con probabilidad q, tiene una media p y una varianza pq. Una distribución B(n,p) puede entenderse como la suma de n binomiales B(1,p), luego aplicando el TCL, si n es grande la distribución B(n,p) se puede aproximar por una normal que tiene como media a np y como varianza npq.

8.2.5. Muestreo

Aunque la teoría que será desarrollada más tarde está referida solamente a muestras aleatorias simples, realizaremos aquí una rápida revisión de posibles métodos para la toma de muestras que podemos encontrarnos en la práctica.

Los pasos a seguir para la recolección de una muestra son los siguientes:

Definir la población en estudio especificando las unidades que la componen, el área geográfica donde se realiza el estudio (si procede) y el periodo de tiempo en el que se realizará el mismo.

- Definir el marco: listado o descripción de los elementos que forman la población.

- Definir la unidad de muestreo: Ciudades, calles, hogares, individuos, etc ...

- Definir las variables a medir o las preguntas que se harán si se trata de una encuesta.

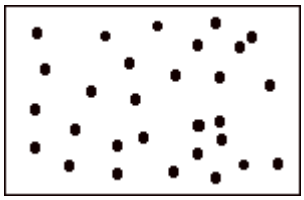
- Seleccionar el método de muestreo: Probabilístico o No Probabilístico, aunque son los primeros los que nos permiten la estimación correcta de parámetros.

- Calcular el tamaño necesario para obtener une determinada precisión en la estimación. Este punto se verá con más detalle en el apartado dedicado a la estimación por intervalos.

- Elaborar el plan de muestreo que guiará el trabajo de campo.

En cuanto al tipo de muestreo, algunas de las características más importantes de los muestreos probabilísticos más usuales se detallan a continuación:

- ✓ **Muestreo aleatorio simple (Mas)**



Se trata de un procedimiento de muestreo (sin reemplazamiento), en el que se seleccionan *n* unidades de las N en la población, de forma que cualquier posible muestra del mismo tamaño tiene la misma probabilidad de ser elegidas.

Se realizan n selecciones independientes de forma que en cada selección los individuos que no han sido elegidos tengan la misma probabilidad de serlo.

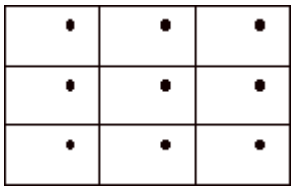
El procedimiento habitual consiste en numerar todos los elementos de la población y se seleccionan muestras del tamaño deseado utilizando una tabla de números aleatorios o un programa de ordenador que proporcione números aleatorios.

Recuérdese que "al azar" no significa "de cualquier manera", para que el procedimiento de muestreo sea válido es necesario utilizar correctamente el proceso de generación de números aleatorios.

Entre las ventajas de este procedimiento esta la compensación de valores altos y bajos con lo que la muestra tiene una composición similar a la de la población, es además un procedimiento sencillo y produce estimadores de los parámetros desconocidos próximos a los valores reales de los mismos.

El principal inconveniente de este tipo de muestreo es que necesita un marco adecuado y amplio que no siempre es fácil de conseguir y que no contiene información a priori sobre la población que podría ser útil en la descripción de la misma.

- ✓ **Muestreo sistemático**



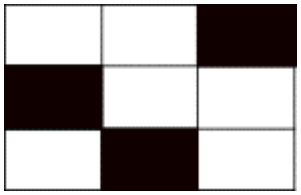
- Se ordenan los individuos de la población y se numeran. - Se divide la población en tantos grupos como individuos se quieren tener en la muestra. Se selecciona uno al azar en el primer grupo y se elige el que ocupa el mismo lugar en todos los grupos.

-La ventaja principal es que es más sencillo y más barato que el muestreo aleatorio simple, además, se comporta igual si no hay patrones o periodicidades en los datos.

-La aparición de patrones desconocidos puede llevar a importantes errores en la estimación de los parámetros.

Este tipo de muestreo puede utilizarse, por ejemplo, en encuestas telefónicas programadas mediante ordenador.

- ✓ **Muestreo por conglomerados**

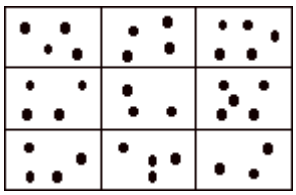


-Se divide la población en grupos de acuerdo con su proximidad geográfica o de otro tipo. (conglomerados). Cada grupo ha de ser heterogéneo y tener representados todos las características de la población.

Por ejemplo, los conglomerados en un estudio sobre la situación de las mujeres en una determinada zona rural pueden ser los municipios de la zona.

-Se selecciona una muestra de conglomerados al azar y se toma el conglomerado completo o una muestra del mismo.
-Necesitan menos información previa sobre los individuos particulares.
-Soluciona el problema de los patrones en los datos.
-Si el número de bloques no es muy grande se puede incurrir en errores de estimación si se han incluido conglomerados atípicos.
-Los conglomerados que se realizan teniendo en cuenta proximidad geográfica pueden no tener un significado importante en la población (no responden a una característica real).
- Este tipo de muestreo se utiliza fundamentalmente para reducir los costes de toma de muestras al tomar grupos de individuos completos.

- ✓ **Muestreo estratificado**



Se divide la población en grupos homogéneos (estratos) de acuerdo con las características a estudiar. Por ejemplo, en un estudio de las características socioeconómicas de una ciudad los estratos pueden ser los barrios de la misma, ya que los barrios suelen presentar características diferenciales.

-Se selecciona una muestra aleatoria de cada estrato tratando de que todos los estratos de la población queden representados.

-Permite utilizar información a priori sobre la estructura de la población en relación con las variables a estudiar.

-Obtiene representantes de todos los estratos de la población.

-Diferentes opciones de selección del tamaño de la muestra en los estratos:

-El mismo número en cada estrato.

-Proporcional. (La más común)

-Optima.

Notas:

-El problema más importante en la realización de una investigación por muestreo es encontrar el marco adecuado (Lista de los elementos de la población que pueden ser incluidos en la muestra).

-En algunos casos es necesario encontrar una población identificable mayor que la población de interés y que incluya a la misma. Por ejemplo, si queremos realizar una encuesta sobre los trabajadores de la construcción de la ciudad de Salamanca y no disponemos de una lista de los mismos, podemos tomar una lista de los cabezas de familias trabajadores o de las viviendas ocupadas. El único problema adicional es que la encuesta será más cara.

8.1. distribuciones muestrales

Todo lo que veremos a continuación está pensado para poblaciones infinitas (muy grandes) y con muestreo aleatorio simple. El muestreo aleatorio simple garantiza una muestra representativa de la población y la obtención de observaciones independientes.

Dada una población X, el proceso de muestreo consiste en obtener, al azar, un valor de la variable X, x1; El valor obtenido puede ser cualquiera de los de la población, luego los posibles valores para x1 son todos los de X, y por tanto x1 puede considerarse como una realización particular (observación) de una variable aleatoria X1 con la misma distribución que X.

A continuación obtenemos, independientemente de la primera observación, un valor x2 que puede considerarse como una realización particular de una variable aleatoria X2 con la misma distribución que X e independiente de X1. Obsérvese que la población no se modifica al extraer uno de sus individuos ya que es infinita. (Si la población es finita podría utilizarse un muestreo con reemplazamiento).

El proceso continúa hasta obtener una muestra de tamaño n, n observaciones x1, x2, ... , xn de n variables aleatorias X1, X2, ... , Xn independientes e idénticamente distribuidas.

Definición: Sea X una variable aleatoria con f.d.p F, y sean X1, X2, ... , Xn , n variables aleatorias independientes con la misma f.d.p F que X. Se dice que X1, X2, ... , Xn , son una muestra aleatoria de tamaño n de F o bien n observaciones independientes de X.

Hemos utilizado letras minúsculas, como en descriptiva, para denotar las observaciones particulares de una muestra, y letras mayúsculas para denotar las variables aleatorias de las que se han tomado. A lo largo de la exposición teórica ambas serán intercambiables y serán utilizadas indistintamente para representar a las correspondientes variables aleatorias.

Otra forma de ver la muestra es como una variable aleatoria multivariante con función de densidad de probabilidad es el producto de las funciones de densidad de cada una de las componentes (ya que son independientes)

$$f(X1, X2, \ldots , Xn) = f(X1)\, f(X2) \ldots f(Xn)$$

Donde las funciones de densidad son iguales a la de X. Esta forma de entender la muestra supera el ámbito de un curso introductorio.

Una vez obtenida la muestra la describimos en términos de algunas de sus características fundamentales como la media, la desviación típica, etc ... A tales características las solemos denominar estadísticos.

Definición: Un estadístico es una función de los valores muestrales que no depende de ningún parámetro poblacional desconocido.

Un estadístico es también una variable aleatoria ya que es una función de variables aleatorias. Por ejemplo la media muestral

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

es una variable aleatoria de la que tenemos una sola observación

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Veámoslo con un ejemplo sencillo

Supongamos que disponemos de una población finita en la que disponemos de 4 individuos que toman los valores {1, 2, 3, 4}.

Supongamos que obtenemos una muestra sin reemplazamiento de tamaño 2. Las distintas posibilidades son

{1, 2} {1, 3} {1, 4} {2, 3} {2, 4} {3, 4}

Obtendremos, dependiendo de la muestra elegida, las siguientes medias respectivamente:

1.5 2 2.5 2.5 3 3.5

Es claro que la media muestral no es un valor fijo sino que puede considerarse también como una variable aleatoria de la que tenemos una sola observación, la media de la muestra concreta seleccionada.

Dicha variable tendrá una distribución de probabilidad asociada. (En este caso una distribución discreta que toma los valores 1.5, 2, 2.5, 3 y 3.5 con probabilidades 1/6, 1/6, 2/6, 1/6, 1/6, respectivamente.

Definición: A la distribución de un estadístico calculado a partir de los valores tomados de una muestra se la denomina distribución muestral del estadístico.

En la mayor parte de los casos supondremos que nuestra población tiene distribución normal y que los estadísticos que vamos a utilizar son la media y la desviación típica (o la cuasi desviación típica).

8.2. distribuciones muestrales de medias

Sea $X_1, X_2, \ldots, X_n$, una muestra aleatoria de una población X en la que $E(X) = \mu$ $Var(X)= \sigma^2$

Entonces el valor esperado (media) y la varianza del estadístico "media muestral" son

$$E(\overline{X}) = \mu$$

$$Var(\overline{X}) = \frac{\sigma^2}{n}$$

$$Desv(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

La comprobación del resultado es obvia si tenemos en cuenta que la esperanza de la suma de varias variables aleatorias independientes es la suma de las esperanzas, y que la varianza es la suma de las varianzas, y además que si multiplicamos una variable por una constante, la varianza queda multiplicada por la constante al cuadrado. Entonces

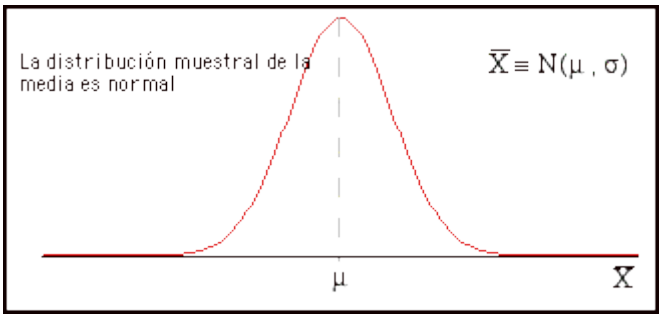
$$E(\overline{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\, n\,\mu = \mu$$

$$Var(\overline{X}) = Var\left(\sum_{i=1}^{n} \frac{X_i}{n}\right) = \sum_{i=1}^{n} \frac{1}{n^2} Var(X_i) = n\frac{\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Si además, la población es normal, es decir, $X \equiv N(\mu, \sigma)$ entonces la media muestral es también normal $\overline{X} \equiv N(\mu, \sigma)$.

Basta tener en cuenta las propiedades de la normal que ya se vieron en su momento.

El resultado es importante en estimación ya que, aunque la media poblacional y la media muestral no coincidan, los posibles valores de la media muestral se concentran de forma simétrica alrededor de la media poblacional, además, la dispersión es menor a medida que aumenta el tamaño muestral.



Distribución muestral de las medias.

La distribución muestral asociada a varianzas y cuasivarianzas es un poco más compleja y su obtención supera los objetivos del curso, de forma que nos limitaremos a exponerlas.

Sea $X_1, X_2, \ldots, X_n$, una muestra aleatoria simple de una población $X \equiv N(\mu, \sigma^2)$, entonces la variable aleatoria

$$\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2}$$

sigue una ji-cuadrado con n-1 grados de libertad.

Del resultado anterior se deduce que las variables

$$\frac{n\,S^2}{\sigma^2} \quad y \quad \frac{(n-1)\,\hat{S}^2}{\sigma^2}$$

donde siguen ambas una ji-cuadrado con n-1 grados de libertad.

8.3. Distribución muestral de s al cuadrado

Otro caso especial muy importante de la distribución gamma se obtiene haciendo $\alpha = \frac{v}{2}\ y\ \beta = 2$ donde v es un entero positivo. El resultado se llama **distribución ji cuadrada**. La distribución tiene un parámetro sencillo, v, que recibe el nombre de **grados de libertad**.

8.2.6. Definiciones:

- ✓ **Distribución ji cuadrada** La variable aleatoria continua X tiene una distribución ji cuadrada, con v grados de libertad, si su función de densidad es:

$$f(x) = \frac{1}{2^{\frac{v}{2}}\Gamma(v/2)}\ x^{\frac{v}{2}-1}e^{-x/2}$$

Para x>0

- ✓ **Distribución ji cuadrada (x^2).** Es la distribución de una variable aleatoria que siempre es positiva, con una posición oblicua hacia la derecha y unimodal. La forma de la distribución depende de un parámetro llamado grados de libertad. La figura 1.1 muestra una distribución ji cuadrada típica.

Estadística de prueba para ji cuadrada. Una variable aleatoria cuya distribución de muestreo es aproximada a la de x2 es:

$$x^2 = \sum \frac{(observada - esperada)^2}{esperada}$$

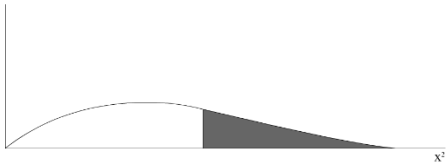


Figura 1.1 Una distribución x2

En realidad la **distribución ji-cuadrada** es la distribución muestral de s^2. O sea que si se extraen todas las muestras posibles de una población normal y a cada muestra se le calcula su varianza, se obtendrá la distribución muestral de varianzas.

Para estimar la varianza poblacional o la desviación estándar, se necesita conocer el estadístico X2. Si se elige una muestra de tamaño n de una población normal con varianza σ^2:

$$\frac{(n-1)s^2}{\sigma^2}$$

El estadístico tiene una distribución muestral que es una distribución ji-cuadrada con gl=n-1 grados de libertad y se denota X^2 (X es la minúscula de la letra griega ji). El estadístico ji-cuadrada está dado por:

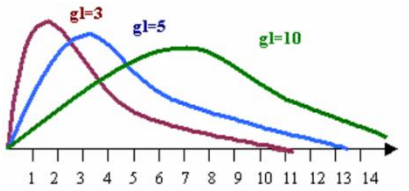
$$X^2 = \frac{(n-1)s^2}{\sigma^2}$$

Donde n es el tamaño de la muestra, s^2 la varianza muestral y σ^2 la varianza de la población de donde se extrajo la muestra. El estadístico ji-cuadrada también se puede dar con la siguiente expresión:

$$X^2 = \frac{\sum (x-\bar{x})^2}{\sigma^2}$$

Propiedades

- ✓ Los valores de X^2 son mayores o iguales que 0.
- ✓ La forma de una distribución X^2 depende del gl=n-1. En consecuencia, hay un número infinito de distribuciones X^2.
- ✓ El área bajo una curva ji-cuadrada y sobre el eje horizontal es 1.
- ✓ Las distribuciones X^2 no son simétricas. Tienen colas estrechas que se extienden a la derecha; esto es, están sesgadas a la derecha.
- ✓ Cuando n>2, la media de una distribución X^2 es n-1 y la varianza es 2(n-l).
- ✓ El valor modal de una distribución X^2 se da en el valor (n-3).
- ✓ La siguiente figura ilustra tres distribuciones X^2.



8.4. distribución T

En las unidades anteriores se manejó el uso de la distribución z, la cual se podía utilizar siempre y cuando los tamaños de las muestras fueran mayores o iguales a 30 ó en muestras más pequeñas si la distribución o las distribuciones de donde proviene la muestra o las muestras son normales.

En esta unidad se podrán utilizar muestras pequeñas siempre y cuando la distribución de donde proviene la muestra tenga un comportamiento normal. Esta es una condición para utilizar las tres distribuciones que se manejarán en esta unidad; t de student, X2 ji-cuadrada y Fisher.

A la teoría de pequeñas muestras también se le llama teoría exacta del muestreo, ya que también la podemos utilizar con muestras aleatorias de tamaño grande.

En esta unidad se verá un nuevo concepto necesario para poder utilizar a las tres distribuciones mencionadas. Este concepto es "grados de libertad".

Para definir grados de libertad se hará referencia a la varianza muestral:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$$

Esta fórmula está basada en *n-1* ***grados de libertad*** (degrees of freedom). Esta terminología resulta del hecho de que si bien s^2 está basada en *n* cantidades $x_1 - \bar{x},\ x_2 - \bar{x},\ \ldots,\ x_n - \bar{x},$ éstas suman cero, así que especificar los valores de cualquier *n-1* de las cantidades determina el valor restante. Por ejemplo, si *n=4 y*

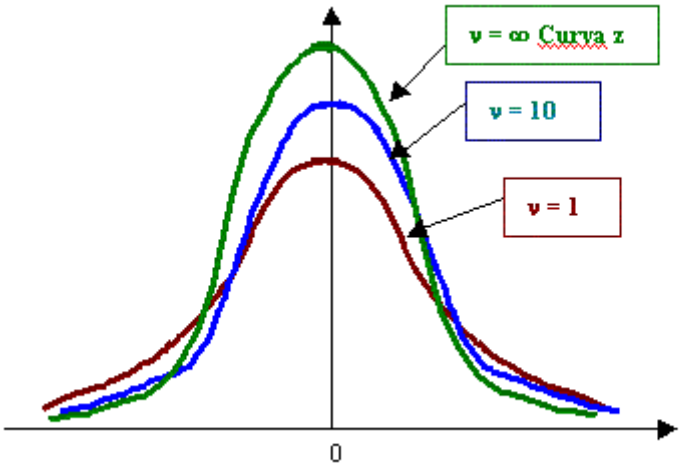
$x_1 - \bar{x} = 8$; $x_2 - \bar{x} = -6$ y $x_4 - \bar{x} = -4$, entonces automáticamente tenemos $x_3 - \bar{x} = 2$, así que sólo tres de los cuatro valores de $x_i - \bar{x}$ están libremente determinamos 3 grados de libertad.

Entonces, en esta unidad la fórmula de grados de libertad será *n-1* y su simbología $\nu = nu.$

Si $\bar{x}$ es el promedio de las *n* observaciones que contiene la muestra aleatoria, entonces la distribución $z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ es una distribución normal estándar. Supóngase que la varianza de la población σ^2 es desconocida. ¿Qué sucede con la distribución de esta estadística si se reemplaza σ por s? La distribución *t* proporciona la respuesta a esta pregunta.

La media y la varianza de la distribución *t* son $\mu = 0$ y $\sigma^2 = \nu/(\nu - 2)$ para $\nu > 2$, respectivamente.

La siguiente figura presenta la gráfica de varias distribuciones *t*. La apariencia general de la distribución *t* es similar a la de la distribución normal estándar: ambas son simétricas y unimodales, y el valor máximo de la ordenada se alcanza en la media $\mu = 0$. Sin embargo, la distribución *t* tiene colas más amplias que la normal; esto es, la probabilidad de las colas es mayor que en la distribución normal. A medida que el número de grados de libertad tiende a infinito, la forma límite de la distribución *t* es la distribución normal estándar.



Propiedades de las distribuciones t

- ✓ Cada curva t tiene forma de campana con centro en 0.
- ✓ Cada curva t, está más dispersa que la curva normal estándar z.
- ✓ A medida que ν aumenta, la dispersión de la curva t correspondiente disminuye.

- ✓ A medida que $v \to \infty$, la secuencia de curvas t se aproxima a la curva normal estándar, por lo que la curva z recibe a veces el nombre de curva t con gl = ∞

La distribución de la variable aleatoria t está dada por:

$$h(t) = \frac{\Gamma[(v+1)/2]}{\Gamma(v/2)\sqrt{\pi v}}\left(1+\frac{t^2}{v}\right)^{-(v+1)/2}, -\infty < t < \infty.$$

Esta se conoce como la distribución t con v grados de libertad.

Sean $X_1, X_2, \ldots, X_n$ variables aleatorias independientes que son todas normales con media μ y desviación estándar σ. Entonces la variable aleatoria $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$ tiene una distribución t con v= n-1 grados de libertad.

La distribución de probabilidad de t se publicó por primera vez en 1908 en un artículo de W. S. Gosset. En esa época, Gosset era empleado de una cervecería irlandesa que desaprobaba la publicación de investigaciones de sus empleados. Para evadir esta prohibición, publicó su trabajo en secreto bajo el nombre de "Student". En consecuencia, la distribución t normalmente se llama distribución *t de Student*, o simplemente distribución t. Para derivar la ecuación de esta distribución, Gosset supone que las muestras se seleccionan de una población normal. Aunque esto parecería una suposición muy restrictiva, se puede mostrar que las poblaciones no normales que poseen distribuciones en forma casi de campana aún proporcionan valores de t que se aproximan muy de cerca a la distribución t.

La distribución t difiere de la de Z en que la varianza de t depende del tamaño de la muestra y siempre es mayor a uno. Unicamente cuando el tamaño de la muestra tiende a infinito las dos distribuciones serán las mismas.

Se acostumbra representar con t_α el valor t por arriba del cual se encuentra un área igual a α. Como la distribución t es simétrica alrededor de una media de cero, tenemos $t_{1-\alpha} = -t_\alpha$; es decir, el valor t que deja un área de $1-\alpha$ a la derecha y por tanto un área de α a la izquierda, es igual al valor t negativo que deja un área de α en la cola derecha de la distribución. Esto es, $t_{0.95} = -t_{0.05}$, $t_{0.99}=-t_{0.01}$, etc.

Para encontrar los valores de t se utilizará la tabla de valores críticos de la distribución t del libro Probabilidad y Estadística para Ingenieros de los autores Walpole, Myers y Myers.

8.5. distribución F

Usada en teoría de probabilidad y estadística, la distribución F es una distribución de probabilidad continua. También se le conoce como distribución F de Snedecor (por George Snedecor) o como distribución F de Fisher-Snedecor (por Ronald Fisher).

Una variable aleatoria de distribución *F* se construye como el siguiente cociente:

$$F = \frac{U_1/d_1}{U_2/d_2}$$

donde

- ✓ U1 y U2 siguen una distribución chi-cuadrado con d1 y d2 grados de libertad respectivamente, y
- ✓ U1 y U2 son estadísticamente independientes.

La distribución F aparece frecuentemente como la distribución nula de una prueba estadística, especialmente en el análisis de varianza. Véase el test F.

La función de densidad de una F(d1, d2) viene dada por

$$g(x) = \frac{1}{\mathrm{B}(d_1/2, d_2/2)} \left(\frac{d_1\, x}{d_1\, x + d_2}\right)^{d_1/2} \left(1 - \frac{d_1\, x}{d_1\, x + d_2}\right)^{d_2/2} x^{-1}$$

para todo número real x ≥ 0, donde d1 y d2 son enteros positivos, y B es la función beta.

La función de distribución es

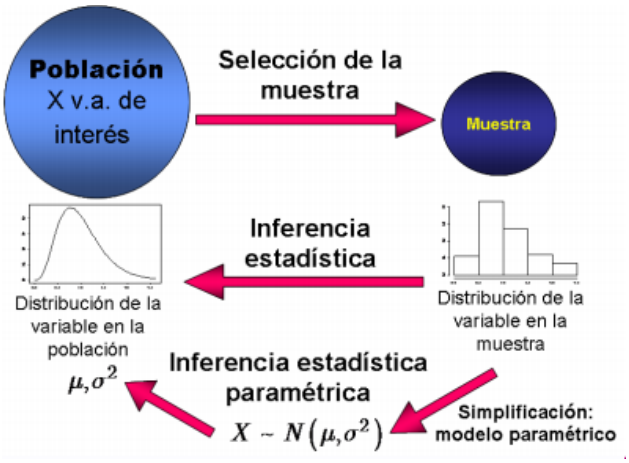
$$G(x) = I_{\frac{d_1 x}{d_1 x + d_2}}(d_1/2, d_2/2)$$

donde I es la función beta incompleta regularizada.

9. PROBLEMAS DE ESTIMACIÓN DE UNA Y DOS MUESTRAS

9.1. introducción

La estimación puntual trata el problema de estimar mediante un número el valor de una característica poblacional o parámetro θ desconocido (por ejemplo, la estimación del IPC de un determinado período). En muchos casos la estimación puntual no es suficiente en el sentido de que no nos indica el error que se comete en dicha estimación. Lo razonable en la práctica es adjuntar, junto a la estimación puntual del parámetro, un cierto intervalo numérico que mida el margen de error que, de acuerdo a las observaciones muestrales, pueda tener dicha estimación. La idea de Intervalo de Confianza es proponer un rango de valores entre los que posiblemente se encuentre el verdadero valor del parámetro θ.



9.2. Inferencia estadística

En inferencia estadística se llama estimación al conjunto de técnicas que permiten dar un valor aproximado de un parámetro de una población a partir de los datos proporcionados por

una muestra. Por ejemplo, una estimación de la media de una determinada característica de una población de tamaño N podría ser la media de esa misma característica para una muestra de tamaño n.1

9.3. Métodos clásicos de estimación

La estimación se divide en tres grandes bloques, cada uno de los cuales tiene distintos métodos que se usan en función de las características y propósitos del estudio:

- ✓ Estimación puntual
  - Método de los momentos;
  - Método de la máxima verosimilitud;
  - Método de los mínimos cuadrados;
- ✓ Estimación por intervalos.
- ✓ Estimación bayesiana.

9.3.1. Estimación puntual

La estimación de parámetros tiene por finalidad asignar valores a los parámetros poblacionales a partir de los estadísticos obtenidos en las muestras. Dicho de otra manera, la finalidad de la estimación de parámetros es caracterizar las poblaciones a partir de la información de las muestras (por ejemplo, inferir el valor de la Media de la población a partir de los datos de la muestra).

La estimación puntual consiste en atribuir **un** valor (la **estimación**) al parámetro poblacional. Si la muestra es representativa de la población, podemos esperar que los estadísticos calculados en las muestras tengan valores semejantes a los parámetros poblacionales, y la estimación consiste en asignar los valores de los estadísticos muestrales a los parámetros poblacionales. Los estadísticos con que obtenemos las estimaciones se denominan **estimadores.**

Ejemplo

Se desea estimar la Media de las puntuaciones del curso 2003/4, pero solo se dispone de 50 puntuaciones seleccionadas aleatoriamente. La Media de la muestra (el estimador), es igual a 5.6 y atribuimos este valor (la estimación) a la Media del curso completo.

Resumiendo:

Media de la población (curso 2003/4) μ= Desconocida.

Estimador: Media de la muestra: $\overline{X}$

Estimación de μ: 5.6

Podemos utilizar como estimadores de la Media de la población otros estadísticos de tendencia central como la Moda o la Mediana, pero NO todos los estimadores son apropiados. Los estimadores deben satisfacer ciertos requisitos, y por esta razón, interesa conocer sus propiedades a fin de utilizar los que sean adecuados según las circunstancias de la estimación.

- ✓ **Características estimadores**

Sesgo. Se dice que un estimador es **insesgado** si la Media de la distribución del estimador es igual al parámetro.

Estimadores insesgados son la Media muestral (estimador de la Media de la población) y la Varianza (estimador de la Varianza de la población):

$$\boxed{\begin{array}{l} \overline{X} \Rightarrow \mu \\ \tilde{s}_X^2 \Rightarrow \sigma^2 \end{array}}$$

Ejemplo

En una población de 500 puntuaciones cuya Media (m) es igual a 5.09 han hecho un muestreo aleatorio (número de muestras= 10000, tamaño de las muestras= 100) y hallan que la Media de las Medias muestrales es igual a **5.09,** (la media poblacional y la media de las medias muestrales coinciden). En cambio, la Mediana de la población es igual a **5** y la Media de las Medianas es igual a **5.1** esto es, hay diferencia ya que la Mediana es un estimador sesgado.

La Varianza es un estimador sesgado. Ejemplo: La Media de las Varianzas obtenidas con la Varianza

$$\boxed{s^2 = \frac{\sum \left(X - \overline{X}(\ )\right)^2}{n}}$$

en un muestreo de 1000 muestras (n=25) en que la Varianza de la población es igual a **9.56** ha resultado igual a **9.12**, esto es, no coinciden. En cambio, al utilizar la Cuasivarianza

$$\boxed{\tilde{s}^2 = \frac{\sum \left(X - \overline{X}\right)^2}{n-1}}$$

la Media de las Varianzas muestrales es igual a 9.5, esto es, coincide con la Varianza de la población ya que la Cuasivarianza es un estimador insesgado.

2) Consistencia. Un estimador es consistente si aproxima el valor del parámetro cuanto mayor es **n** (tamaño de la muestra).

Algunos estimadores consistentes son:

$$\boxed{\begin{array}{l} \overline{X} \Rightarrow \mu \\ \tilde{s}_X^2 \Rightarrow \sigma^2 \end{array}}$$

Ejemplo

En una población de 500 puntuaciones cuya Media (m) es igual a **4.9** han hecho tres muestreos aleatorios (número de muestras= 100) con los siguientes resultados:

n	Media de las Medias muestrales
5	4.6
25	4.8
100	4.9

vemos que el muestreo en que n=100 la Media de las Medias muestrales toma el mismo valor que la Media de la población.

3) Eficiencia. Diremos que un estimador es más eficiente que otro si la Varianza de la distribución muestral del estimador es menor a la del otro estimador. Cuanto menor es la eficiencia, menor es la confianza de que el estadístico obtenido en la muestra aproxime al parámetro poblacional.

Ejemplo

La Varianza de la distribución muestral de la Media en un muestreo aleatorio (número de muestras: 1000, n=25) ha resultado igual a **0.4.** La Varianza de la distribución de Medianas ha resultado, en el mismo muestreo, igual a **1.12**, (este resultado muestra que la Media es un estimador más eficiente que la Mediana).

9.3.2. Estimación por intervalos

La estimación por intervalos consiste en establecer el intervalo de valores donde es más probable se encuentre el parámetro. La obtención del intervalo se basa en las siguientes consideraciones:

a) Si conocemos la distribución muestral del estimador podemos obtener las probabilidades de ocurrencia de los estadísticos muestrales.

b) Si **conociéramos** el valor del parámetro poblacional, **podríamos** establecer la probabilidad de que el estimador se halle dentro de los intervalos de la distribución muestral.

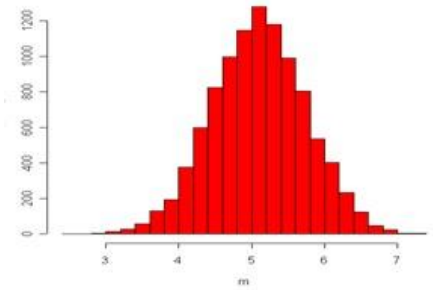
c) El problema es que el parámetro poblacional es desconocido, y por ello el intervalo se establece alrededor del estimador. Si repetimos el muestreo un gran número de veces y definimos un intervalo alrededor de cada valor del estadístico muestral, el parámetro se sitúa dentro de cada intervalo en un porcentaje conocido de ocasiones. Este intervalo es denominado "**intervalo de confianza**".

Ejemplo

Se generan 100000 muestras aleatorias (n=25) de una población que sigue la distribución Normal, y resulta:

	Población	Distribución muestral
Media	5.1	5.1
Desviación Típica	3.2	0.6

La distribución de las Medias muestrales aproxima al modelo Normal:



En consecuencia, el intervalo dentro del cual se halla el 95% de las Medias muestrales es

$$\boxed{\mu_{\bar{X}} \pm 1.96\sigma_{\bar{X}} = 5.1 \pm (1.96)(0.6) = \begin{cases} 6.3 \\ 3.9 \end{cases}}$$

(Nota: Los valores +-1.96 que multiplican la Desviación Típica de la distribución muestral son los valores cuya función de distribución es igual a 0.975 y 0.025 respectivamente y se pueden obtener en las tablas de la distribución Normal estandarizada o de funciones en aplicaciones informáticas como Excel). Seguidamente generamos una muestra de la población y obtenemos su Media, que es igual a 4.5. Si establecemos el intervalo alrededor de la Media muestral, el parámetro poblacional (5.1) está incluido dentro de sus límites:

$$\boxed{\bar{X} \pm 1.96\sigma_{\bar{X}} = 4.5 \pm (1.96)(0.6) = \begin{cases} 5.7 \\ 3.3 \end{cases}}$$

Ahora bien, la distancia de un punto A a un punto B es la misma que de B a A. Por esa razón, la distancia desde m a la Media muestral es la misma que va de la Media muestral a m. En consecuencia, si hacemos un muestreo con un número grande de muestras observamos que el 95% de las veces (aproximadamente) el valor de la Media de la población (m) se encuentra dentro del intervalo definido alrededor de **cada uno** de los valores de la Media muestral. El porcentaje de veces que el valor de m se halla dentro de **alguno** de los intervalos de confianza es del 95%, y es denominado **nivel de confianza**.

Si queremos establecer un intervalo de confianza en que el % de veces que m se halle dentro del intervalo sea igual al 99%, la expresión anterior es:

$$\boxed{\bar{X} \pm 2.58\sigma_{\bar{X}}}$$

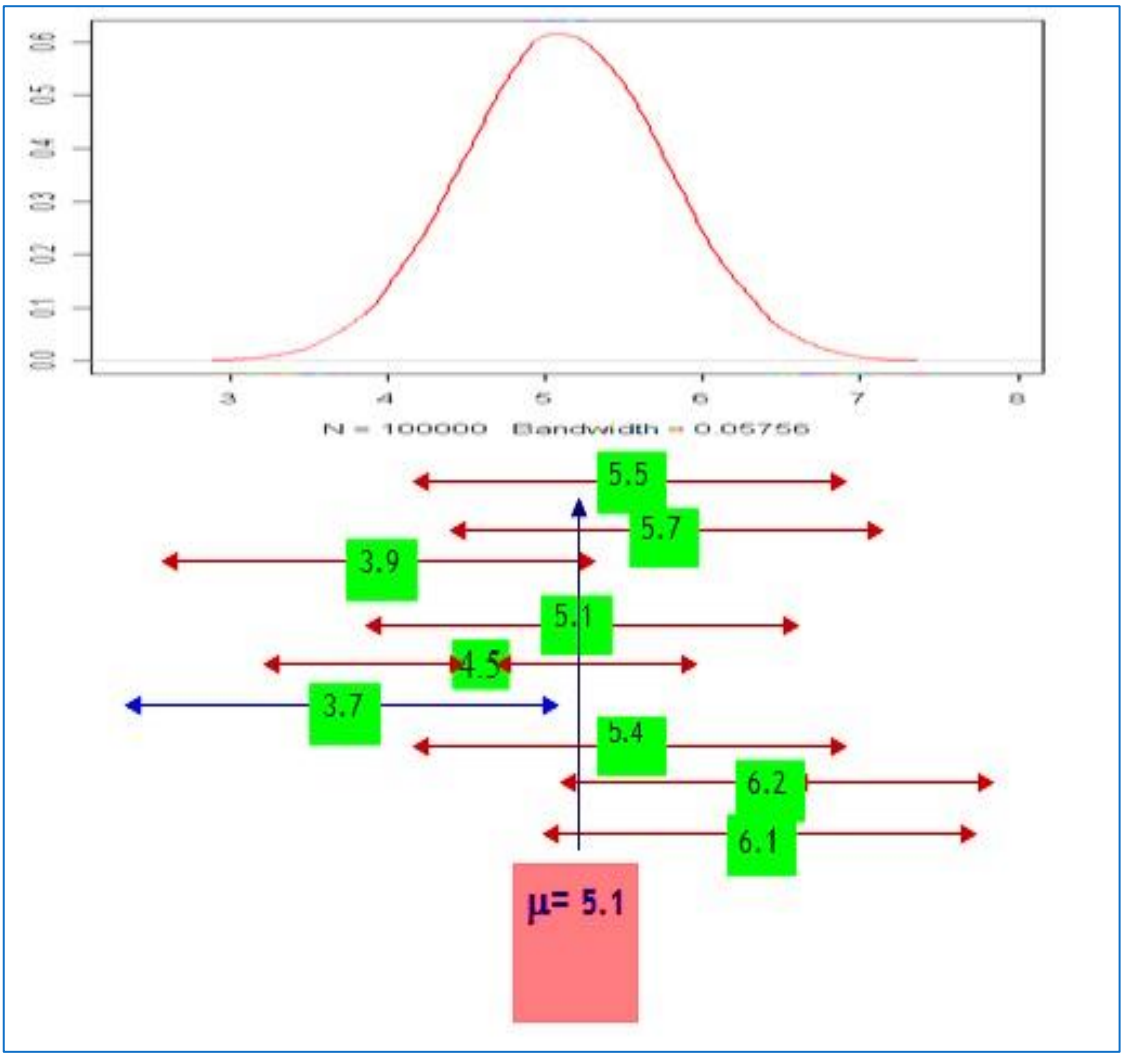
(Obtenemos el valor +-2.58 que multiplica la Desviación Típica de la distribución muestral en las tablas de la distribución Normal estandarizada o de funciones en aplicaciones informáticas como Excel), y son los valores cuya función de probabilidad es igual a 0.995 y 0.005 respectivamente).

Ejemplo

La siguiente imagen muestra la distribución de las Medias muestrales obtenidas de 100000 muestras aleatorias y los intervalos alrededor de cada una de las Medias obtenidas de diez de las muestras:

$$\overline{X} \pm 1.96\sigma_{\overline{X}} = \begin{cases} l_s \\ l_i \end{cases}$$

donde l_s y l_e simbolizan los límites superior e inferior del intervalo de confianza al 95%.



Nueve de los diez intervalos (salvo el definido alrededor de la Media muestral igual a 3.7) incluyen el valor del parámetro dentro sus límites.

9.4. Una sola muestra estimación de la media

La distribución muestral de $\bar{X}$ está centrada en μ y en la mayoría de las aplicaciones la varianza es más pequeña que la de cualesquiera otros estimadores de μ. Por lo tanto, se utilizará la media muestral $\bar{x}$ como una estimación puntual para la media de la población μ. Recuerde que $\sigma^2_{\bar{X}} = \sigma^2/n$, por lo que una muestra grande producirá un valor de $\bar{X}$ procedente de una distribución muestral con varianza pequeña. Por consiguiente, es probable que $\bar{x}$ sea una estimación muy precisa de μ cuando n es grande.

Consideremos ahora la estimación por intervalos de μ. Si seleccionamos nuestra muestra a partir de una población normal o, a falta de ésta, si n es suficientemente grande, podemos establecer un intervalo de confianza para μ considerando la distribución muestral de $\bar{X}$.

De acuerdo con el teorema del límite central, podemos esperar que la distribución muestral de $\bar{X}$ esté distribuida de forma aproximadamente normal con media $\mu_{\bar{X}} = \mu$ y desviación estándar $\sigma_{\bar{X}} = \sigma/\sqrt{n}$. Al escribir $z_{\alpha/2}$ para el valor z por arriba del cual encontramos una área de $\alpha/2$ bajo la curva normal, en la figura 9.2 podemos ver que

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha,$$

donde

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

En consecuencia,

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha.$$

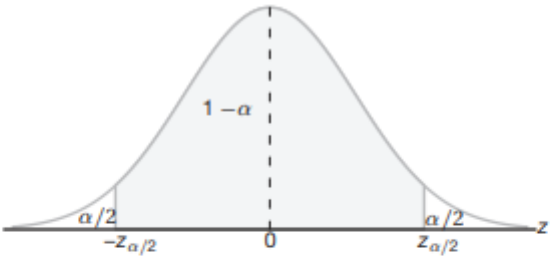


Figura 9.2: $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$.

Si multiplicamos cada término en la desigualdad por $\sigma/\sqrt{n}$ y después restamos $\bar{X}$ de cada término, y en seguida multiplicamos por − 1 (para invertir el sentido de las desigualdades), obtenemos

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

Se selecciona una muestra aleatoria de tamaño n de una población cuya varianza σ^2 se conoce y se calcula la media $\bar{x}$ para obtener el intervalo de confianza $100(1-\alpha)\%$. Es importante enfatizar que recurrimos al teorema del límite central citado anteriormente. Como resultado, es importante observar las condiciones para las aplicaciones que siguen.

Intervalo de confianza de μ cuando se conoce σ^2

Si $\bar{x}$ es la media de una muestra aleatoria de tamaño n de una población de la que se conoce su varianza σ^2, lo que da un intervalo de confianza de $100(1-\alpha)\%$ para μ es

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$

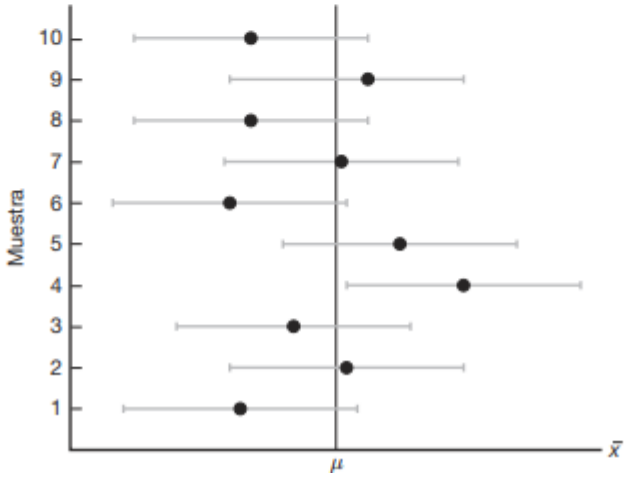
donde $z_{\alpha/2}$ es el valor z que deja una área de $\alpha/2$ a la derecha.

En el caso de muestras pequeñas que se seleccionan de poblaciones no normales, no podemos esperar que nuestro grado de confianza sea preciso. Sin embargo, para muestras de tamaño $n \geq 30$, en las que la forma de las distribuciones no esté muy sesgada, la teoría de muestreo garantiza buenos resultados.

Queda claro que los valores de las variables aleatorias $\hat{\Theta}_L$ y $\hat{\Theta}_U$, las cuales se definieron en la sección 9.3, son los límites de confianza

$$\hat{\theta}_L = \bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \text{y} \quad \hat{\theta}_U = \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Muestras diferentes producirán valores diferentes de $\bar{x}$ y, por lo tanto, producirán diferentes estimaciones por intervalos del parámetro μ, como se muestra en la figura 9.3. Los puntos en el centro de cada intervalo indican la posición de la estimación puntual $\bar{x}$ para cada muestra aleatoria. Observe que todos los intervalos tienen el mismo ancho, pues esto depende sólo de la elección de $z_{\alpha/2}$ una vez que se determina $\bar{x}$. Cuanto más grande sea el valor de $z_{\alpha/2}$ que elijamos, más anchos haremos todos los intervalos, y podremos tener más confianza en que la muestra particular que seleccionemos producirá un intervalo que contenga el parámetro desconocido μ. En general, para una elección de $z_{\alpha/2}$, $100(1-\alpha)\%$ de los intervalos contendrá μ.



Límites de confianza unilaterales

Los intervalos de confianza y los límites de confianza resultantes que hasta ahora hemos analizado en realidad son *bilaterales,* es decir, tienen límites superior e inferior. Sin embargo, hay muchas aplicaciones en las que sólo se requiere un límite. Por ejemplo, si a un ingeniero le interesara determinar una medida de resistencia a la tensión, la información que más le ayudaría a lograr su objetivo sería la del límite inferior, ya que éste indica el escenario del "peor caso", es decir, el de la menor resistencia. Por otro lado, si se buscara determinar una medida para la cual un valor de μ relativamente grande no fuera redituable o deseable, entonces la medida que resultaría de interés sería la del límite de confianza superior. Un ejemplo en el que la medida del límite superior sería muy informativa es el caso en el que se necesita hacer inferencias para determinar la composición media de mercurio en el agua de un río.

Los límites de confianza unilaterales se desarrollan de la misma forma que los intervalos bilaterales. Sin embargo, la fuente es un enunciado de probabilidad unilateral que utiliza el teorema del límite central:

$$P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_\alpha\right) = 1 - \alpha.$$

Entonces, es posible manipular el enunciado de probabilidad de forma muy similar a como se hizo anteriormente para obtener

$$P(\mu > \bar{X} - z_\alpha \sigma/\sqrt{n}) = 1 - \alpha.$$

Una manipulación similar de $P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > -z_\alpha\right) = 1 - \alpha$ da

$$P(\mu < \bar{X} + z_\alpha \sigma/\sqrt{n}) = 1 - \alpha.$$

Como resultado, se obtienen los siguientes límites unilaterales superior e inferior.

Límites de confianza unilaterales de μ cuando se conoce el valor de σ^2

Si $\bar{X}$ es la media de una muestra aleatoria de tamaño n a partir de una población con varianza σ^2, los límites de confianza unilaterales del $100(1 - \alpha)\%$ para μ son dados por

límite unilateral superior: $\bar{x} + z_\alpha \sigma/\sqrt{n}$;

límite unilateral inferior: $\bar{x} - z_\alpha \sigma/\sqrt{n}$.

El caso en que se desconoce σ

Con frecuencia debemos tratar de estimar la media de una población sin conocer la varianza. El lector debería recordar que en el capítulo 8 aprendió que, si tenemos una muestra aleatoria a partir de una *distribución normal,* entonces la variable aleatoria

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

tiene una distribución t de Student con $n - 1$ grados de libertad. Aquí S es la desviación estándar de la muestra. En esta situación, en la que se desconoce σ, se puede utilizar T para construir un intervalo de confianza para μ. El procedimiento es igual que cuando se conoce σ, sólo que en este caso σ se reemplaza con S y la distribución normal estándar se reemplaza con la distribución t. Si nos remitimos a la figura 9.5, podemos afirmar que

$$P(-t_{\alpha/2} < T < t_{\alpha/2}) = 1 - \alpha,$$

donde $t_{\alpha/2}$ es el valor t con $n - 1$ grados de libertad, por arriba del cual encontramos una área de $\alpha/2$. Debido a la simetría, un área igual de $\alpha/2$ caerá a la izquierda de $-t_{\alpha/2}$. Al sustituir por T escribimos

$$P\left(-t_{\alpha/2} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2}\right) = 1 - \alpha.$$

Al multiplicar cada término en la desigualdad por $S/\sqrt{n}$ y después restar $\bar{X}$ de cada término y multiplicar por -1, obtenemos

$$P\left(\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right) = 1 - \alpha.$$

Para nuestra muestra aleatoria particular de tamaño n se calculan la media $\bar{x}$ y la desviación estándar s, y se obtiene el siguiente intervalo de confianza $100(1-\alpha)\%$ para μ.

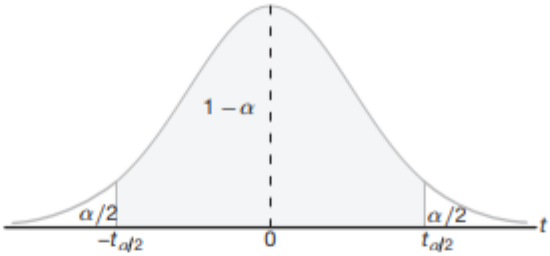


Figura 9.5: $P(-t_{\alpha/2} < T < t_{\alpha/2}) = 1 - \alpha$.

Intervalo de confianza para μ cuando se desconoce σ^2

Si $\bar{x}$ y s son la media y la desviación estándar de una muestra aleatoria de una población normal de la que se desconoce la varianza σ^2, un intervalo de confianza del $100(1-\alpha)\%$ para μ es

$$\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}},$$

donde $t_{\alpha/2}$ es el valor t con $v = n - 1$ grados de libertad que deja una área de $\alpha/2$ a la derecha.

Hicimos una distinción entre los casos en los que se conoce σ y en los que se desconoce calculando las estimaciones del intervalo de confianza. Deberíamos resaltar que para el caso en que se conoce σ se utiliza el teorema del límite central, mientras que, para el caso en que se desconoce, se usa la distribución muestral de la variable aleatoria T. Sin embargo, el uso de la distribución t se basa en la premisa de que el muestreo es de una distribución normal. Siempre que la forma de la distribución se aproxime a la de campana, se puede utilizar la distribución t para calcular los intervalos de confianza cuando se desconoce σ^2, y se pueden esperar muy buenos resultados.

Los límites de confianza unilaterales calculados para μ con σ desconocida son como el lector esperaría, a saber:

$$\bar{x} + t_{\alpha}\frac{s}{\sqrt{n}} \qquad \text{y} \qquad x - t_{\alpha}\frac{s}{\sqrt{n}}.$$

Éstos son, respectivamente, los límites superior e inferior del $100(1-\alpha)\%$. Aquí t_{α} es el valor t que tiene una área α a la derecha.

Concepto de intervalo de confianza para una muestra grande

Con frecuencia los estadísticos recomiendan que incluso cuando no sea posible suponer la normalidad, se desconozca σ y $n \geq 30$, σ se puede reemplazar con s para poder utilizar el intervalo de confianza

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

A menudo se hace referencia a esto como un *intervalo de confianza para una muestra grande*. La justificación para esto reside sólo en la presunción de que, con una muestra tan grande como 30 y una distribución de la población no muy sesgada, s estará muy cerca de la σ verdadera y, de esta manera, el teorema del límite central continuará siendo válido. Se debería destacar que esto es sólo una aproximación y que la calidad de los resultados mejora a medida que aumenta el tamaño de la muestra.

9.5. Error estándar de una estimación puntual

Hicimos una distinción muy clara entre los objetivos de las estimaciones puntuales y las estimaciones del intervalo de confianza. Las primeras proporcionan un solo número que se extrae de un conjunto de datos experimentales, y las segundas proporcionan un intervalo razonable para el parámetro, *dados los datos experimentales*; es decir, $100(1-\alpha)\%$ de tales intervalos que se calcula "cubren" el parámetro.

Estos dos métodos de estimación se relacionan entre sí. El elemento en común es la distribución muestral del estimador puntual. Considere, por ejemplo, el estimador $\bar{X}$ de μ cuando se conoce σ. Indicamos antes que una medida de la calidad de un estimador insesgado es su varianza. La varianza de $\bar{X}$ es

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}.$$

De esta forma, la desviación estándar de $\bar{X}$ o *error estándar de* $\bar{X}$ es $\sigma/\sqrt{n}$. En términos simples, el error estándar de un estimador es su desviación estándar. Para el caso de $\bar{X}$ el límite de confianza que se calcula

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \text{ se escribe como } \bar{x} \pm z_{\alpha/2}\ \text{e.e.}(\bar{x}),$$

donde "e.e." es el error estándar. El punto importante es que el ancho del intervalo de confianza de μ depende de la calidad del estimador puntual a través de su error estándar. En el caso en que se desconoce σ y la muestra proviene de una distribución normal, s reemplaza a σ y se incluye el *error estándar estimado* $S/\sqrt{n}$. Por consiguiente, los límites de confianza de μ son:

Límites de confianza para μ cuando se desconoce σ^2

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} = \bar{x} \pm t_{\alpha/2}\ \text{e.e.}(\bar{x})$$

De nuevo, el intervalo de confianza *no es mejor* (en términos de anchura) *que la calidad de la estimación puntual,* en este caso a través de su error estándar estimado. A menudo el software de computación se refiere a los errores estándar estimados simplemente como "errores estándar".

A medida que avanzamos a intervalos de confianza más complejos, prevalece el concepto de que el ancho de los intervalos de confianza se acorta cuando mejora la calidad de la estimación puntual correspondiente, aunque no siempre es tan sencillo como aquí se ilustra. Se puede argumentar que un intervalo de confianza es tan sólo una ampliación de la estimación puntual para tomar en cuenta la exactitud de dicha estimación.

9.6. Intervalos de tolerancia

La estimación puntual y la estimación por intervalos de la media que se expusieron en las secciones 9.4 y 9.5 proporcionan buena información del parámetro desconocido μ de una distribución normal, o de una distribución no normal a partir de la cual se toma una muestra grande. Algunas veces, además de la media de la población, el experimentador podría estar interesado en predecir el **valor posible de una observación futura**. Por ejemplo, en el control de calidad el experimentador podría necesitar utilizar los datos observados para predecir una nueva observación. Un proceso de manufactura de una pieza de metal se podría evaluar basándose en si la pieza cumple con las especificaciones de resistencia a la tensión. En ciertas ocasiones un cliente podría estar interesado en comprar una **sola pieza**. En este caso un intervalo de confianza de la resistencia media a la tensión no cubriría la información requerida. El cliente necesitaría una aseveración respecto a la incertidumbre de una **sola observación**. Este tipo de requerimiento se satisface muy bien construyendo un **intervalo de predicción**.

Es muy sencillo obtener un intervalo de predicción para las situaciones que hemos considerado hasta el momento. Suponga que la muestra aleatoria se tomó de una población normal con media μ desconocida y varianza σ^2 conocida. Un estimador puntual natural de una nueva observación es $\bar{X}$. En la sección 8.4 se aprendió que la varianza de $\bar{X}$ es σ^2/n. Sin embargo, para predecir una nueva observación no basta con explicar la variación debida a la estimación de la media, también tendríamos que explicar la **variación de una observación futura**. A partir de la suposición sabemos que la varianza del error aleatorio en una nueva observación es σ^2. El desarrollo de un intervalo de predicción se representa mejor empezando con una variable aleatoria normal $x_0 - \bar{x}$, donde x_0 es la nueva observación y $\bar{x}$ se toma de la muestra. Como x_0 y $\bar{x}$ son independientes, sabemos que

$$z = \frac{x_0 - \bar{x}}{\sqrt{\sigma^2 + \sigma^2/n}} = \frac{x_0 - \bar{x}}{\sigma\sqrt{1 + 1/n}}$$

es $n(z; 0, 1)$. Como resultado, si utilizamos el enunciado de probabilidad

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$

con el estadístico z anterior, y si colocamos x_0 en el centro del enunciado de probabilidad, tenemos que la probabilidad de que ocurra el siguiente evento es $1 - \alpha$:

$$\bar{x} - z_{\alpha/2}\sigma\sqrt{1 + 1/n} < x_0 < \bar{x} + z_{\alpha/2}\sigma\sqrt{1 + 1/n}.$$

Como resultado, el intervalo de predicción calculado se formaliza como sigue.

Intervalo de predicción para una observación futura cuando se conoce σ^2

Para una distribución normal de mediciones con media μ desconocida y varianza σ^2 conocida, un **intervalo de predicción** del $100(1 - \alpha)\%$ de una observación futura x_0 es

$$\bar{x} - z_{\alpha/2}\sigma\sqrt{1 + 1/n} < x_0 < \bar{x} + z_{\alpha/2}\sigma\sqrt{1 + 1/n},$$

donde $z_{\alpha/2}$ es el valor z que deja una área de $\alpha/2$ a la derecha.

Uso de límites de predicción para detectar valores extremos

Hasta el momento hemos puesto poca atención al concepto de **valores extremos** u observaciones aberrantes. La mayoría de los investigadores científicos son muy sensibles a la existencia de observaciones de valores extremos, también llamados datos defectuosos o "malos". En el capítulo 12 profundizaremos en el estudio de este concepto. Sin embargo, nos interesa considerarlos aquí porque la detección de los valores extremos está estrechamente relacionada con los intervalos de predicción.

Para nuestros propósitos nos conviene considerar que una observación extrema es una que proviene de una población con una media diferente a la que determina el resto de la muestra de tamaño n que se está estudiando. El intervalo de predicción produce un límite que "cubre" una sola observación futura con probabilidad $1 - \alpha$, si ésta proviene de la población de la que se tomó la muestra. Por lo tanto, una metodología para detectar valores extremos implica la regla de que **una observación es un valor extremo si cae fuera del intervalo de predicción calculado sin incluir la observación cuestionable en la muestra**. Como resultado, para el intervalo de predicción del ejemplo 9.8, en el caso de los paquetes de carne, la observación que se obtiene al medir un nuevo paquete y encontrar que su contenido libre de grasa está fuera del intervalo (93.96, 98.44) se podría considerar como un valor extremo.

9.7. Límites de tolerancia

Como vimos en la sección 9.6, el científico o el ingeniero podrían estar menos interesados en estimar parámetros que en obtener información sobre el lugar en el que caería una *observación* o medición individual. Este tipo de situaciones requiere intervalos de predicción. Sin embargo, existe un tercer tipo de intervalo que es útil en muchas aplicaciones. Una vez más, suponga que el interés se centra en torno a la fabricación de la pieza de un componente y que existen especificaciones sobre una dimensión de esa parte. Además, la media de esa dimensión no es tan importante. Sin embargo, a diferencia del escenario de la sección 9.6, se podría estar menos interesado en una sola observación y más en el lugar en el que cae la mayoría de la población. Si las especificaciones del proceso son importantes, el administrador del proceso se interesará en el desempeño a largo plazo, **no en la siguiente observación**. Debemos tratar de determinar los límites que, en cierto sentido probabilístico, "cubren" los valores en la población, es decir, los valores medidos de la dimensión.

Un método para establecer el límite deseado consiste en determinar un intervalo de confianza sobre una *proporción fija* de las mediciones. Esto se comprende mejor visualizando una situación en la que se realiza un muestreo aleatorio de una distribución normal con media conocida μ y varianza σ^2. Evidentemente, un límite que cubre el 95% central de la población de observaciones es

$$\mu \pm 1.96\sigma.$$

A esto se le llama **intervalo de tolerancia** y, en realidad, su cobertura del 95% de las observaciones medidas es exacta. Sin embargo, en la práctica rara vez se conocen μ y σ; por consiguiente, el usuario debe aplicar

$$\bar{x} \pm ks.$$

Ahora bien, el intervalo es, desde luego, una variable aleatoria, por lo tanto, la *cobertura* de una proporción de la población por el intervalo no es exacta. Como resultado, se debe usar un intervalo de confianza del $100(1 - \gamma)\%$, ya que no se puede esperar que $\bar{x} \pm ks$ cubra cualquier proporción específica todo el tiempo. Lo anterior nos lleva a la siguiente definición.

Límites de tolerancia

Para una distribución normal de mediciones en la que se desconoce la media μ y la desviación estándar σ, los **límites de tolerancia** son dados por $\bar{x} \pm ks$, donde k se determina de tal manera que se pueda estar seguro, con un $100(1-\gamma)\%$ de confianza, de que los límites dados contienen al menos la proporción $1-\alpha$ de las mediciones.

La tabla A.7 ofrece valores de k para $1-\alpha = 0.90, 0.95, 0.99$; $\gamma = 0.05, 0.01$; y para valores seleccionados de n de 2 a 300.

Diferencia entre intervalos de confianza, intervalos de predicción e intervalos de tolerancia

Es importante resaltar la diferencia entre los tres tipos de intervalos que se estudiaron e ilustraron en las secciones anteriores. Los cálculos son sencillos, pero la interpretación podría resultar confusa. En aplicaciones de la vida real tales intervalos no son intercambiables, ya que sus interpretaciones son muy diferentes.

En el caso de los intervalos de confianza sólo se pone atención en la **media de la población**. Por ejemplo, el ejercicio 9.13 de la página 283 se refiere a un proceso de ingeniería que produce alfileres para costura. Se establece una especificación sobre la dureza de Rockwell por debajo de la cual el cliente no aceptará ningún alfiler. En este caso un parámetro de la población debe tener poca relevancia. Es importante que el ingeniero sepa en dónde *van a estar la mayoría de los valores de la dureza de Rockwell*. Por consiguiente, se deberían utilizar los límites de tolerancia. Seguramente, al administrador le agradará saber que los límites de tolerancia en cualquier producto del proceso son más rigurosos que las especificaciones para el propio proceso.

Es verdad que la interpretación del límite de tolerancia se relaciona hasta cierto punto con el intervalo de confianza. El intervalo de tolerancia del $100(1-\alpha)\%$ sobre, digamos, la proporción 0.95, se podría considerar como un intervalo de confianza **sobre el 95% intermedio** de la distribución normal correspondiente. Los límites de tolerancia unilaterales también son relevantes. En el caso del problema de dureza de Rockwell se desearía tener un límite inferior de la forma $\bar{x} - ks$, tal que se tenga un 99% de confianza en que al menos 99% de los valores de la dureza de Rockwell excederán al valor calculado.

Los intervalos de predicción se pueden aplicar cuando es importante determinar un límite para un **solo valor**. Aquí la media no es la cuestión, ni tampoco la ubicación de la mayoría de la población, lo que se requiere, más bien, es la ubicación de una sola nueva observación.

9.8. Dos muestras estimación de la diferencia entre 2 medias

Si tenemos dos poblaciones con medias μ_1 y μ_2, y varianzas σ_1^2 y σ_2^2, respectivamente, el estadístico que da un estimador puntual de la diferencia entre μ_1 y μ_2 es $\bar{X}_1 - \bar{X}_2$. Por lo tanto, para obtener una estimación puntual de $\mu_1 - \mu_2$, se seleccionan dos muestras aleatorias independientes, una de cada población, de tamaños n_1 y n_2, y se calcula $\bar{x}_1 - \bar{x}_2$, la diferencia de las medias muestrales. Evidentemente, debemos considerar la distribución muestral de $\bar{X}_1 - \bar{X}_2$.

De acuerdo con el teorema 8.3, podemos esperar que la distribución muestral de $\bar{X}_1 - \bar{X}_2$ esté distribuida de forma aproximadamente normal con media $\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$ y desviación estándar $\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$. Por lo tanto, podemos asegurar, con una probabilidad de $1-\alpha$, que la variable normal estándar

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

caerá entre $-z_{\alpha/2}$ y $z_{\alpha/2}$. Si nos remitimos una vez más a la figura 9.2, escribimos

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha.$$

Al sustituir para Z, establecemos de manera equivalente que

$$P\left(-z_{\alpha/2} < \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} < z_{\alpha/2}\right) = 1 - \alpha,$$

que conduce al siguiente intervalo de confianza del $100(1-\alpha)\%$ para $\mu_1 - \mu_2$.

Intervalo de confianza para $\mu_1 - \mu_2$ cuando se conocen σ_1^2 y σ_2^2

Si $\bar{x}_1$ y $\bar{x}_2$ son las medias de muestras aleatorias independientes de tamaños n_1 y n_2, de poblaciones que tienen varianzas conocidas σ_1^2 y σ_2^2, respectivamente, un intervalo de confianza del $100(1-\alpha)\%$ para $\mu_1 - \mu_2$ es dado por

$$(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},$$

donde $z_{\alpha/2}$ es el valor z que deja una área de $\alpha/2$ a la derecha.

El grado de confianza es exacto cuando las muestras se seleccionan de poblaciones normales. Para poblaciones no normales el teorema del límite central permite una buena aproximación para muestras de tamaño razonable.

9.9. Observaciones pareadas

Ahora estudiaremos los procedimientos de estimación para la diferencia de dos medias cuando las muestras no son independientes y las varianzas de las dos poblaciones no son necesariamente iguales. La situación que se considera aquí tiene que ver con una condición experimental muy especial, a saber, *las observaciones pareadas*. A diferencia de la situación que se describió antes, las condiciones de las dos poblaciones no se asignan de forma aleatoria a las unidades experimentales. Más bien, cada unidad experimental homogénea recibe ambas condiciones de la población; como resultado, cada unidad experimental tiene un par de observaciones, una para cada población. Por ejemplo, si realizamos una prueba de una nueva dieta con 15 individuos, los pesos antes y después de seguir la dieta conforman la información de las dos muestras. Las dos poblaciones son "antes" y "después", y la unidad experimental es el individuo. Evidentemente, las observaciones en un par tienen algo en común. Para determinar si la dieta es efectiva consideramos las diferencias $d_1, d_2, \ldots, d_n$ en las observaciones pareadas. Estas diferencias son los valores de una muestra aleatoria $D_1, D_2, \ldots, D_n$ de una población de diferencias, que supondremos distribuidas normalmente, con media $\mu_D = \mu_1 - \mu_2$ y varianza σ_D^2. Estimamos σ_D^2 mediante s_d^2, la varianza de las diferencias que constituyen nuestra muestra. El estimador puntual de μ_D es dado por $\bar{D}$.

¿Cuándo debe hacerse el pareado?

Parear observaciones en un experimento es una estrategia que se puede emplear en muchos campos de aplicación. Se expondrá al lector a tal concepto en el material relacionado con la prueba de hipótesis en el capítulo 10 y en los temas de diseño experimental en los capítulos 13 y 15. Al seleccionar unidades experimentales relativamente homogéneas (dentro de las unidades) y permitir que cada unidad experimente ambas condiciones de la población, se reduce la varianza del error experimental efectiva (en este caso σ_D^2). El lector puede visualizar la i-ésima diferencia del par como

$$D_i = X_{1i} - X_{2i}.$$

la prueba de hipótesis en el capítulo 10 y en los temas de diseño experimental en los capítulos 13 y 15. Al seleccionar unidades experimentales relativamente homogéneas (dentro de las unidades) y permitir que cada unidad experimente ambas condiciones de la población, se reduce la varianza del error experimental efectiva (en este caso σ_D^2). El lector puede visualizar la i-ésima diferencia del par como

$$D_i = X_{1i} - X_{2i}.$$

Como las dos observaciones se toman de la unidad experimental de la muestra no son independientes y, de hecho,

$$\text{Var}(D_i) = \text{Var}(X_{1i} - X_{2i}) = \sigma_1^2 + \sigma_2^2 - 2\,\text{Cov}(X_{1i}, X_{2i}).$$

Entonces, de manera intuitiva, se espera que σ_D^2 debería reducirse debido a la similitud en la naturaleza de los "errores" de las dos observaciones dentro de una unidad experimental, a lo cual se llega mediante la expresión anterior. En realidad se espera que, si la unidad es homogénea, la covarianza sea positiva. Como resultado, la ganancia en calidad del intervalo de confianza sobre la que se obtuvo sin parear es mayor cuando hay homogeneidad dentro de las unidades y cuando las diferencias grandes van de una a otra unidad. Se debería tener en cuenta que el desempeño del intervalo de confianza dependerá del error estándar de $\bar{D}$, que es, por supuesto, $\sigma_D/\sqrt{n}$, donde n es el número de pares. Como indicamos antes, la intención al parear es reducir σ_D.

Equilibrio entre reducir la varianza y perder grados de libertad

Al comparar los intervalos de confianza obtenidos con y sin pareado es evidente que hay un intercambio implicado. Aunque en realidad el pareado debería reducir la varianza y, por lo tanto, el error estándar de la estimación puntual, los grados de libertad disminuyen al reducir el problema a uno con una sola muestra. Como resultado, el punto $t_{\alpha/2}$ ligado al error estándar se ajusta en concordancia. De esta manera, el pareado podría resultar contraproducente. Esto ocurriría con certeza si se experimenta sólo una reducción modesta en la varianza (a través de σ_D^2) mediante el pareado.

Otra ilustración del pareado implicaría elegir n pares de sujetos, donde cada par tenga una característica similar, como el coeficiente intelectual (CI), la edad o la raza, y luego para cada par seleccionar un miembro al azar para obtener un valor de X_1, dejando que el otro miembro proporcione el valor de X_2. En este caso, X_1 y X_2 podrían representar las calificaciones obtenidas por dos individuos con igual CI cuando uno es asignado al azar a un grupo que usa el método de enseñanza convencional y al otro a un grupo que utiliza materiales programados.

Se puede establecer un intervalo de confianza del $100(1-\alpha)\%$ para μ_D escribiendo

$$P(-t_{\alpha/2} < T < t_{\alpha/2}) = 1 - \alpha,$$

donde $T = \frac{\bar{D} - \mu_D}{S_d/\sqrt{n}}$ y $t_{\alpha/2}$, como antes, es un valor de la distribución t con $n-1$ grados de libertad.

En la actualidad se acostumbra reemplazar T por su definición en la desigualdad anterior y desarrollar los pasos matemáticos que conduzcan al siguiente intervalo de confianza del $100(1-\alpha)\%$ para $\mu_1 - \mu_2 = \mu_D$.

Intervalo de confianza para $\mu_D = \mu_1 - \mu_2$ para observaciones pareadas

Si $\bar{d}$ y s_d son la media y la desviación estándar, respectivamente, de las diferencias distribuidas normalmente de n pares aleatorios de mediciones, un intervalo de confianza del $100(1-\alpha)\%$ para $\mu_D = \mu_1 - \mu_2$ es

$$\bar{d} - t_{\alpha/2}\frac{s_d}{\sqrt{n}} < \mu_D < \bar{d} + t_{\alpha/2}\frac{s_d}{\sqrt{n}},$$

donde $t_{\alpha/2}$ es el valor t con $v = n - 1$ grados de libertad, que deja una área de $\alpha/2$ a la derecha.

9.10. Una sola muestra estimación de una proporción

El estadístico $\hat{P} = X/n$, en donde X representa el número de éxitos en n ensayos, provee un estimador puntual de la proporción p en un experimento binomial. Por lo tanto, la proporción de la muestra $\hat{p} = x/n$ se utilizará como el estimador puntual del parámetro p.

Si no se espera que la proporción p desconocida esté demasiado cerca de 0 o de 1, se puede establecer un intervalo de confianza para p considerando la distribución muestral de $\hat{P}$. Si en cada ensayo binomial asignamos el valor 0 a un fracaso y el valor 1 a un éxito, el número de éxitos, x, se puede interpretar como la suma de n valores que consta sólo de ceros y unos, y $\hat{p}$ es sólo la media muestral de esos n valores. En consecuencia, por el teorema del límite central, para n suficientemente grande $\hat{P}$ está distribuida de forma casi normal con media

$$\mu_{\hat{P}} = E(\hat{P}) = E\left(\frac{X}{n}\right) = \frac{np}{n} = p$$

y varianza

$$\sigma^2_{\hat{P}} = \sigma^2_{X/n} = \frac{\sigma^2_X}{n^2} = \frac{npq}{n^2} = \frac{pq}{n}.$$

Por lo tanto, podemos afirmar que

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha, \text{ con } Z = \frac{\hat{P} - p}{\sqrt{pq/n}},$$

y $z_{\alpha/2}$ es el valor por arriba del cual encontramos una área de $\alpha/2$ debajo de la curva normal estándar. Al sustituir para Z escribimos

$$P\left(-z_{\alpha/2} < \frac{\hat{P} - p}{\sqrt{pq/n}} < z_{\alpha/2}\right) = 1 - \alpha.$$

Cuando n es grande se introduce un error muy pequeño sustituyendo el estimado puntual $\hat{p} = x/n$ para la p debajo del signo de radical. Entonces podemos escribir

$$P\left(\hat{P} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{P} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}\right) \approx 1 - \alpha.$$

Por otro lado, al resolver para p en la desigualdad cuadrática anterior,

$$-z_{\alpha/2} < \frac{\hat{P} - p}{\sqrt{pq/n}} < z_{\alpha/2},$$

obtenemos otra forma del intervalo de confianza para p con los siguientes límites:

$$\frac{\hat{p} + \frac{z^2_{\alpha/2}}{2n}}{1 + \frac{z^2_{\alpha/2}}{n}} \pm \frac{z_{\alpha/2}}{1 + \frac{z^2_{\alpha/2}}{n}}\sqrt{\frac{\hat{p}\hat{q}}{n} + \frac{z^2_{\alpha/2}}{4n^2}}.$$

Para una muestra aleatoria de tamaño n se calcula la proporción muestral $\hat{p} = x/n$ y se pueden obtener los siguientes intervalos de confianza aproximados del $100(1-\alpha)\%$ para p.

Intervalos de confianza para p de una muestra grande

Si $\hat{p}$ es la proporción de éxitos en una muestra aleatoria de tamaño n, y $\hat{q} = 1 - \hat{p}$, un intervalo de confianza aproximado del $100(1-\alpha)\%$ para el parámetro binomial p se obtiene por medio de (método 1)

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

o mediante (método 2)

$$\frac{\hat{p} + \frac{z^2_{\alpha/2}}{2n}}{1 + \frac{z^2_{\alpha/2}}{n}} - \frac{z_{\alpha/2}}{1 + \frac{z^2_{\alpha/2}}{n}}\sqrt{\frac{\hat{p}\hat{q}}{n} + \frac{z^2_{\alpha/2}}{4n^2}} < p < \frac{\hat{p} + \frac{z^2_{\alpha/2}}{2n}}{1 + \frac{z^2_{\alpha/2}}{n}} + \frac{z_{\alpha/2}}{1 + \frac{z^2_{\alpha/2}}{n}}\sqrt{\frac{\hat{p}\hat{q}}{n} + \frac{z^2_{\alpha/2}}{4n^2}},$$

donde $z_{\alpha/2}$ es el valor z que deja una área de $\alpha/2$ a la derecha.

Cuando n es pequeña y se cree que la proporción desconocida p se acerca a 0 o a 1, el procedimiento del intervalo de confianza que se establece aquí no es confiable y, por lo tanto, no se debería emplear. Para estar seguros se requiere que tanto $n\hat{p}$ como $n\hat{q}$ sean mayores que o iguales a 5. Los métodos para calcular un intervalo de confianza para el parámetro binomial p también se pueden aplicar cuando se está utilizando la distribución binomial con el fin de aproximar la distribución hipergeométrica; es decir, cuando n es pequeña respecto a N, como se ilustra en el ejemplo 9.14.

Observe que, aunque el método 2 produce resultados más precisos, su cálculo es más complicado, y la ventaja en precisión que brinda disminuye cuando el tamaño de la muestra es lo suficientemente grande. Debido a esto en la práctica es más común utilizar el método 1.

9.11. Dos muestras estimación de la diferencia entre dos proporciones

Considere el problema en el que se busca estimar la diferencia entre dos parámetros binomiales p_1 y p_2. Por ejemplo, p_1 podría ser la proporción de fumadores con cáncer de pulmón y p_2 la proporción de no fumadores con cáncer de pulmón, y el problema consistiría en estimar la diferencia entre estas dos proporciones. Primero seleccionamos muestras aleatorias independientes de tamaños n_1 y n_2 a partir de las dos poblaciones binomiales con medias n_1p_1 y n_2p_2, y varianzas $n_1p_1q_1$ y $n_2p_2q_2$, respectivamente, después determinamos los números x_1 y x_2 de personas con cáncer de pulmón en cada muestra, y formamos las proporciones $\hat{p}_1 = x_1/n$ y $\hat{p}_2 = x_2/n$. El estadístico $\hat{P}_1 - \hat{P}_2$ provee un estimador puntual de la diferencia entre las dos proporciones, $p_1 - p_2$. Por lo tanto, la diferencia de las proporciones muestrales, $\hat{p}_1 - \hat{p}_2$, se utilizará como la estimación puntual de $p_1 - p_2$.

Se puede establecer un intervalo de confianza para $p_1 - p_2$ considerando la distribución muestral de $\hat{P}_1 - \hat{P}_2$. De la sección 9.10 sabemos que $\hat{P}_1$ y $\hat{P}_2$ están distribuidos cada uno de forma aproximadamente normal, con medias p_1 y p_2, y varianzas p_1q_1/n_1 y p_2q_2/n_2, respectivamente. Al elegir muestras independientes de las dos poblaciones nos aseguramos de que las variables $\hat{P}_1$ y $\hat{P}_2$ serán independientes y luego, por la propiedad reproductiva de la distribución normal que se estableció en el teorema 7.11, concluimos que $\hat{P}_1 - \hat{P}_2$ está distribuido de forma aproximadamente normal con media

$$\mu_{\hat{P}_1 - \hat{P}_2} = p_1 - p_2$$

y varianza

$$\sigma^2_{\hat{P}_1 - \hat{P}_2} = \frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}.$$

Por lo tanto, podemos asegurar que

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha,$$

donde

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\sqrt{p_1 q_1/n_1 + p_2 q_2/n_2}}$$

y $z_{\alpha/2}$ es un valor por arriba del cual encontramos una área de $\alpha/2$ debajo de la curva normal estándar. Al sustituir para Z escribimos

$$P\left[-z_{\alpha/2} < \frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\sqrt{p_1 q_1/n_1 + p_2 q_2/n_2}} < z_{\alpha/2}\right] = 1 - \alpha.$$

Después de realizar las operaciones matemáticas usuales reemplazamos p_1, p_2, q_1 y q_2 bajo el signo de radical por sus estimaciones $\hat{p}_1 = x_1/n_1$, $\hat{p}_2 = x_2/n_2$, $\hat{q}_1 = 1 - \hat{p}_1$ y $\hat{q}_2 = 1 - \hat{p}_2$, siempre y cuando $n_1\hat{p}_1$, $n_2\hat{q}_1$, $n_2\hat{p}_2$ y $n_1\hat{q}_1$ sean todas mayores que o iguales a 5, y se obtiene el siguiente intervalo de confianza aproximado del $100(1-\alpha)\%$ para $p_1 - p_2$.

Intervalo de confianza para $p_1 - p_2$ de una muestra grande

Si $\hat{p}_1$ y $\hat{p}_2$ son las proporciones de éxitos en muestras aleatorias de tamaños n_1 y n_2, respectivamente, $\hat{q}_1 = 1 - \hat{p}_1$ y $\hat{q}_2 = 1 - \hat{p}_2$, un intervalo de confianza aproximado del $100(1-\alpha)\%$ para la diferencia de dos parámetros binomiales $p_1 - p_2$ es dado por

$$(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}},$$

donde $z_{\alpha/2}$ es el valor z que deja una área de $\alpha/2$ a la derecha.

9.12. una sola muestra estimación de la varianza

Si extraemos una muestra de tamaño n de una población normal con varianza σ^2 y calculamos la varianza muestral s^2, obtenemos un valor del estadístico S^2. Esta varianza muestral calculada se utiliza como una estimación puntual de σ^2. En consecuencia, al estadístico S^2 se le denomina estimador de σ^2.

Se puede establecer una estimación por intervalos de σ^2 utilizando el estadístico

$$X^2 = \frac{(n-1)S^2}{\sigma^2}.$$

De acuerdo con el teorema 8.4, cuando las muestras se toman de una población normal el estadístico X^2 tiene una distribución chi cuadrada con $n - 1$ grados de libertad. Podemos escribir (véase la figura 9.7)

$$P(\chi^2_{1-\alpha/2} < X^2 < \chi^2_{\alpha/2}) = 1 - \alpha,$$

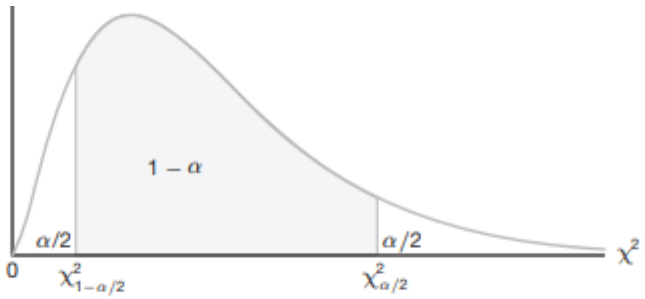


Figura 9.7: $P\left(\chi^2_{1-\alpha/2} < X^2 < \chi^2_{\alpha/2}\right) = 1 - \alpha$.

donde $\chi^2_{1-\alpha/2}$ y $\chi^2_{\alpha/2}$ son valores de la distribución chi cuadrada con $n-1$ grados de libertad, que dejan áreas de $1-\alpha/2$ y $\alpha/2$, respectivamente, a la derecha. Al sustituir para X^2 escribimos

$$P\left[\chi^2_{1-\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2}\right] = 1 - \alpha.$$

Si dividimos cada término de la desigualdad entre $(n-1)S^2$, y después invertimos cada término (lo que cambia el sentido de las desigualdades), obtenemos

$$P\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right] = 1 - \alpha.$$

Para una muestra aleatoria de tamaño n, tomada de una población normal, se calcula la varianza muestral s^2 y se obtiene el siguiente intervalo de confianza del $100(1-\alpha)\%$ para σ^2.

Intervalo de confianza para σ^2

Si s^2 es la varianza de una muestra aleatoria de tamaño n de una población normal, un intervalo de confianza del $100(1-\alpha)\%$ para σ^2 es

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}},$$

donde $\chi^2_{\alpha/2}$ y $\chi^2_{1-\alpha/2}$ son valores χ^2 con $v = n-1$ grados de libertad, que dejan áreas de $\alpha/2$ y $1-\alpha/2$, respectivamente, a la derecha.

Un intervalo de confianza aproximado a $100(1-\alpha)\%$ para σ se obtiene tomando la raíz cuadrada de cada extremo del intervalo para σ^2.

9.13. dos muestras estimación de la proporcion de dos varianzas

Una estimación puntual de la proporción de dos varianzas de la población σ_1^2/σ_2^2 es dada por la proporción s_1^2/s_2^2 de las varianzas muestrales. En consecuencia, el estadístico S_1^2/S_2^2 se conoce como un estimador de σ_1^2/σ_2^2.

Si σ_1^2 y σ_2^2 son las varianzas de poblaciones normales, podemos establecer una estimación por intervalos de σ_1^2/σ_2^2 usando el estadístico

$$F = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}.$$

De acuerdo con el teorema 8.8, la variable aleatoria F tiene una distribución F con $v_1 = n_1 - 1$ y $v_2 = n_2 - 1$ grados de libertad. Por lo tanto, podemos escribir (véase la figura 9.8)

$$P[f_{1-\alpha/2}(v_1, v_2) < F < f_{\alpha/2}(v_1, v_2)] = 1 - \alpha,$$

donde $f_{1-\alpha/2}(v_1, v_2)$ y $f_{\alpha/2}(v_1, v_2)$ son los valores de la distribución F con v_1 y v_2 grados de libertad, que dejan áreas de $1-\alpha/2$ y $\alpha/2$, respectivamente, a la derecha.

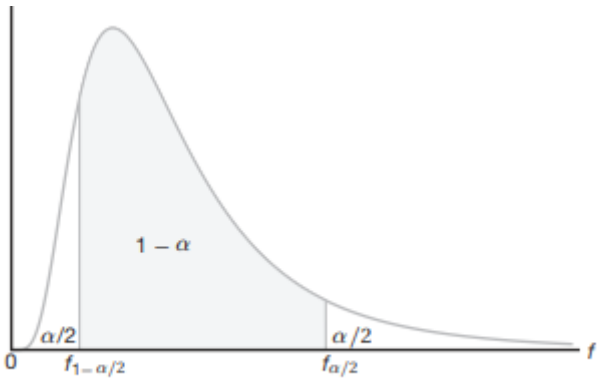


Figura 9.8: $P[f_{1-\alpha/2}(v_1, v_2) < F < f_{\alpha/2}(v_1, v_2)] = 1 - \alpha$.

Al sustituir para F, escribimos

$$P\left[f_{1-\alpha/2}(v_1, v_2) < \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2} < f_{\alpha/2}(v_1, v_2)\right] = 1 - \alpha.$$

Si multiplicamos cada término de la desigualdad por S_2^2/S_1^2, y después invertimos cada término, obtenemos

$$P\left[\frac{S_1^2}{S_2^2}\frac{1}{f_{\alpha/2}(v_1, v_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2}\frac{1}{f_{1-\alpha/2}(v_1, v_2)}\right] = 1 - \alpha.$$

Los resultados del teorema 8.7 nos permiten reemplazar la cantidad $f_{1-\alpha/2}(v_1, v_2)$ por $1/f_{\alpha/2}(v_2, v_1)$. Por lo tanto,

$$P\left[\frac{S_1^2}{S_2^2}\frac{1}{f_{\alpha/2}(v_1, v_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2} f_{\alpha/2}(v_2, v_1)\right] = 1 - \alpha.$$

Para cualesquiera dos muestras aleatorias independientes de tamaño n_1 y n_2 que se seleccionan de dos poblaciones normales, se calcula la proporción de las varianzas muestrales s_1^2/s_2^2 y se obtiene el siguiente intervalo de confianza del $100(1-\alpha)\%$ para σ_1^2/σ_2^2.

Intervalo de confianza para σ_1^2/σ_2^2

Si s_1^2 y s_2^2 son las varianzas de muestras independientes de tamaño n_1 y n_2, respectivamente, tomadas de poblaciones normales, entonces un intervalo de confianza del $100(1-\alpha)\%$ para σ_1^2/σ_2^2 es

$$\frac{s_1^2}{s_2^2}\frac{1}{f_{\alpha/2}(v_1, v_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2} f_{\alpha/2}(v_2, v_1),$$

donde $f_{\alpha/2}(v_1, v_2)$ es un valor f con $v_1 = n_1 - 1$ y $v_2 = n_2 - 1$ grados de libertad que deja una área de $\alpha/2$ a la derecha, y $f_{\alpha/2}(v_2, v_1)$ es un valor f similar con $v_2 = n_2 - 1$ y $v_1 = n_1 - 1$ grados de libertad.

9.14. Estimación de la máxima verosimilitud (opcional)

A menudo los estimadores de parámetros han tenido que recurrir a la intuición. El estimador $\bar{X}$ ciertamente parece razonable como estimador de una media de la población μ. La virtud de S^2 como estimador de σ^2 se destaca en el estudio de estimadores insesgados de la sección 9.3. El estimador para un parámetro binomial p es simplemente una proporción de la muestra que, desde luego, es un *promedio* y recurre al sentido común. Sin embargo, hay muchas situaciones en las que no es del todo evidente cuál debería ser el estimador adecuado. Como resultado, el estudiante de estadística tiene mucho que aprender respecto a las diferentes filosofías que producen distintos métodos de estimación. En esta sección estudiaremos el **método de máxima verosimilitud**.

La estimación por máxima verosimilitud representa uno de los métodos de estimación más importantes en toda la estadística inferencial. No explicaremos el método de manera detallada; más bien, intentaremos transmitir la filosofía de la máxima verosimilitud e ilustrarla con ejemplos que la relacionan con otros problemas de estimación que se examinan en este capítulo.

Función de verosimilitud

Como el nombre lo indica, el método de máxima verosimilitud es aquel para el que se maximiza la *función de verosimilitud*, lo cual se ilustra mejor con un ejemplo que incluye una distribución discreta y un solo parámetro. Consideremos que $X_1, X_2, \ldots, X_n$ son las variables aleatorias independientes tomadas de una distribución de probabilidad discreta representada por $f(\mathbf{x}, \theta)$, donde θ es un solo parámetro de la distribución. Ahora bien,

es la *distribución conjunta de las variables aleatorias,* la cual a menudo se denomina función de probabilidad. Observe que la variable de la función de probabilidad es θ, no **x**. Represente con $x_1, x_2, \ldots, x_n$ los valores observados en una muestra. En el caso de una variable aleatoria discreta, la interpretación es muy clara. La cantidad $L(x_1, x_2, \ldots, x_n; \theta)$, *la verosimilitud de la muestra*, es la siguiente probabilidad conjunta:

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \mid \theta),$$

que es la probabilidad de obtener los valores muestrales $x_1, x_2, \ldots, x_n$. Para el caso discreto el estimador de máxima verosimilitud es el que da como resultado un valor máximo para esta probabilidad conjunta, o el que maximiza la probabilidad de la muestra.

Considere un ejemplo ficticio en el cual se inspeccionan tres artículos que salen de una línea de ensamble. Los artículos se clasifican como defectuosos o no defectuosos, de manera que se aplica el proceso de Bernoulli. La inspección de los tres artículos da como resultado dos artículos no defectuosos seguidos por uno defectuoso. Nos interesa estimar p, la proporción de artículos no defectuosos en el proceso. La probabilidad de la muestra para este ejemplo es dada por

$$p \cdot p \cdot q = p^2 q = p^2 - p^3,$$

donde $q = 1 - p$. La estimación de máxima verosimilitud daría un estimado de p para el que se maximiza la verosimilitud. Resulta claro que si diferenciamos la verosimilitud respecto a p, igualamos la derivada a cero y la resolvemos, obtenemos el valor

$$\hat{p} = \frac{2}{3}.$$

Entonces, desde luego, en esta situación $\hat{p} = 2/3$ es la proporción muestral defectuosa y, por ello, un estimador razonable de la probabilidad de un artículo defectuoso. El lector debería intentar comprender que la filosofía de la estimación de máxima verosimilitud proviene de la noción de que el estimador razonable de un parámetro que se basa en información muestral *es el valor del parámetro que produce la mayor probabilidad de obtener la muestra.* Ésta es, de hecho, la interpretación para el caso discreto, ya que la verosimilitud es la probabilidad de observar de manera conjunta los valores en la muestra.

Definición 9.3: Dadas las observaciones independientes $x_1, x_2, \ldots, x_n$ de una función de densidad de probabilidad (caso continuo) o de una función de masa de probabilidad (caso discreto) $f(\mathbf{x}, \theta)$, el estimador de máxima verosimilitud $\hat{\theta}$ es el que maximiza la función de probabilidad

$$L(x_1, x_2, \ldots, x_n; \theta) = f(\mathbf{x}; \theta) = f(x_1, \theta) f(x_2, \theta) \cdots f(x_n, \theta).$$

Muy a menudo conviene trabajar con el logaritmo natural de la función de verosimilitud para encontrar el máximo de esa función. Considere el siguiente ejemplo acerca del parámetro μ de una distribución de Poisson.

Comentarios adicionales respecto a la estimación de máxima verosimilitud

Un análisis detallado de las propiedades de la estimación de máxima verosimilitud está fuera del alcance de este libro y, por lo general, es un tema importante en un curso teórico de estadística inferencial. El método de máxima verosimilitud permite al analista utilizar el conocimiento de la distribución para determinar un estimador adecuado. *El método de máxima verosimilitud no se puede aplicar si no se conoce la distribución subyacente.* En el ejemplo 9.21 aprendimos que el estimador de máxima verosimilitud no necesariamente carece de sesgo. El estimador de máxima verosimilitud es insesgado *asintóticamente* o *en el límite*; es decir, la magnitud del sesgo se aproxima a cero a medida que la muestra se hace más grande. Al principio de este capítulo examinamos la noción de eficacia, que se vincula con la propiedad de la varianza de un estimador. Los estimadores de máxima verosimilitud tienen propiedades de varianza deseables en el límite. El lector debería consultar la obra de Lehmann y D'Abrera (1998) para más detalles.

9.15. Posibles riesgos y errores conceptuales: relación con el material de otros capítulos

El concepto de *intervalo de confianza de muestra grande* en una población a menudo confunde a los alumnos principiantes. Se basa en la idea de que incluso cuando se desconoce σ y no se está convencido de que la distribución que se muestrea es normal, se puede calcular un intervalo de confianza para μ a partir de

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}.$$

En la práctica es común que se utilice esta fórmula cuando la muestra es demasiado pequeña. El origen de este intervalo de muestra grande es, por supuesto, el teorema del límite central (TLC), con el cual la normalidad no es necesaria. Aquí el TLC requiere una σ conocida, de la cual s sólo es un estimado. Por lo tanto, n debe ser al menos tan grande como 30 y la distribución subyacente debe tener una simetría similar, en cuyo caso el intervalo sigue siendo una aproximación.

Hay casos en que la aplicación práctica del material de este capítulo depende en gran medida del contexto específico. Un ejemplo muy importante es el uso de la distribución t para el intervalo de confianza de μ cuando se desconoce σ. En términos estrictos, el uso de la distribución t requiere que la distribución de donde se toma la muestra sea normal. Sin embargo, es bien sabido que cualquier aplicación de la distribución t es razonablemente insensible, es decir, **robusta**, a la suposición de normalidad. Esto representa una de esas situaciones afortunadas que ocurren con frecuencia en el campo de la estadística, donde no se sostiene un supuesto básico y "¡todo resulta bien!" Sin embargo, la población de la que se toma la muestra no se puede desviar mucho de la normalidad. Por consiguiente, a menudo se recurrirá a las gráficas de probabilidad normal estudiadas en el capítulo 8 y las pruebas de bondad del ajuste que se presentarán en el capítulo 10 para atribuir algún sentido de "cercanía a la normalidad". Esta idea de "robustez a la normalidad" se volverá a presentar en el capítulo 10.

10. PRUEBAS DE HIPÓ TESIS DE UNA Y DOS MUESTRAS

10.1. Hipó tesis estadística

Como se expuso en el capítulo 9, a menudo el problema al que se enfrentan el científico o el ingeniero no es tanto la estimación de un parámetro de la población, sino la formación de un procedimiento de decisión que se base en los datos y que pueda producir una conclusión acerca de algún sistema científico. Por ejemplo, un investigador médico puede decidir con base en evidencia experimental si beber café incrementa el riesgo de cáncer en los seres humanos; un ingeniero quizá tenga que decidir con base en datos muestrales si hay una diferencia entre la precisión de un tipo de medidor y la de otro; o tal vez un sociólogo desee reunir los datos apropiados que le permitan decidir si el tipo de sangre y el color de ojos de un individuo son variables independientes. En cada uno de estos casos el científico o el ingeniero *postulan* o *conjeturan* algo acerca de un sistema. Además, cada uno debe utilizar datos experimentales y tomar decisiones basadas en ellos. En cada caso la conjetura se puede expresar en forma de hipótesis estadística. Los procedimientos que conducen a la aceptación o al rechazo de hipótesis estadísticas como éstas comprenden una área importante de la inferencia estadística. Empecemos por definir con precisión lo que entendemos por **hipótesis estadística**.

Definición 10.1: Una **hipótesis estadística** es una aseveración o conjetura respecto a una o más poblaciones.

La verdad o falsedad de una hipótesis estadística nunca se sabe con absoluta certeza, a menos que se examine toda la población, lo cual, por supuesto, sería poco práctico en la mayoría de las situaciones. En vez de eso se toma una muestra aleatoria de la población de interés y se utilizan los datos contenidos en ella para proporcionar evidencia que respalde o no la hipótesis. La evidencia de la muestra que es inconsistente con la hipótesis planteada conduce al rechazo de la misma.

El papel que desempeña la probabilidad en la prueba de hipótesis

Debería quedar claro al lector que un procedimiento de toma de decisiones debe implicar la conciencia de la *probabilidad de llegar a una conclusión errónea*. Por ejemplo, suponga que la hipótesis que postuló el ingeniero es que la fracción p de artículos defectuosos en cierto proceso es 0.10. El experimento consiste en observar una muestra aleatoria del producto en cuestión. Suponga que se prueban 100 artículos y que se encuentran 12 defectuosos. Es razonable concluir que esta evidencia no rechaza la condición de que el parámetro binomial $p = 0.10$, por lo que puede provocar que no se rechace la hipótesis. Sin embargo, también puede provocar que no se refute $p = 0.12$, o quizá incluso $p = 0.15$. Como resultado, el lector se debe acostumbrar a la idea de que **el rechazo de una hipótesis implica que fue refutada por la evidencia de la muestra**. En otras palabras, **el rechazo significa que existe una pequeña probabilidad de obtener la información muestral observada cuando, de hecho, la hipótesis es verdadera**. Por ejemplo, en la hipótesis de la proporción de artículos defectuosos, una muestra de 100 artículos que revela que hay 20 defectuosos es ciertamente evidencia para el rechazo. ¿Por qué? Si en realidad $p = 0.10$, la probabilidad de obtener 20 o más artículos defectuosos es aproximadamente de 0.002. Con el pequeño riesgo resultante de llegar a una conclusión errónea parecería seguro **rechazar la hipótesis** de que $p = 0.10$. En otras palabras, el rechazo de una hipótesis tiende a casi "descartar" la hipótesis. Por otro lado, es muy importante enfatizar que la aceptación o, más bien, la falta de rechazo no descarta otras posibilidades. Como resultado, *el analista de datos establece una conclusión firme cuando se rechaza una hipótesis.*

La hipótesis nula y la hipótesis alternativa

La estructura de la prueba de hipótesis se establece usando el término **hipótesis nula**, el cual se refiere a cualquier hipótesis que se desea probar y se denota con H_0. El rechazo de H_0 conduce a la aceptación de una **hipótesis alternativa**, que se denota con H_1. La comprensión de las diferentes funciones que desempeñan la hipótesis nula (H_0) y la hipótesis alternativa (H_1) es fundamental para entender los principios de la prueba de hipótesis. La hipótesis alternativa H_1 por lo general representa la *pregunta que se responderá o la teoría que se probará*, por lo que su especificación es muy importante. La hipótesis nula H_0 *anula o se opone a* H_1 y a menudo es el complemento lógico de H_1. A medida que el lector aprenda más sobre la prueba de hipótesis notará que el analista llega a una de las siguientes dos conclusiones:

rechazar H_0 a favor de H_1 debido a evidencia suficiente en los datos o
no rechazar H_0 debido a evidencia insuficiente en los datos.

Observe que las *conclusiones no implican una "aceptación de* H_0*" formal y literal.* La aseveración de H_0 a menudo representa el "status quo" contrario a una nueva idea, conjetura, etcétera, enunciada en H_1; en tanto que no rechazar H_0 representa la conclusión adecuada. En nuestro ejemplo binomial la cuestión práctica podría ser el interés en que la probabilidad histórica de artículos defectuosos de 0.10 ya no sea verdadera. De hecho, la conjetura podría ser que p excede a 0.10. Entonces podríamos afirmar que

$$H_0: p = 0.10,$$
$$H_1: p > 0.10.$$

Ahora, 12 artículos defectuosos de cada 100 no refutan $p = 0.10$, por lo que la conclusión es "no rechazar H_0". Sin embargo, si los datos revelan 20 artículos defectuosos de cada 100, la conclusión sería "rechazar H_0" a favor de H_1: $p > 0.10$.

Aunque las aplicaciones de la prueba de hipótesis son muy abundantes en trabajos científicos y de ingeniería, quizás el mejor ejemplo para un principiante sea el dilema que enfrenta el jurado en un juicio. Las hipótesis nula y alternativa son

H_0: el acusado es inocente,
H_1: el acusado es culpable.

La acusación proviene de una sospecha de culpabilidad. La hipótesis H_0 (el status quo) se establece en oposición a H_1 y se mantiene a menos que se respalde H_1 con evidencia "más allá de una duda razonable". Sin embargo, en este caso "no rechazar H_0" no implica inocencia, sino sólo que la evidencia fue insuficiente para lograr una condena. Por lo tanto, el jurado no necesariamente *acepta* H_0 sino que *no rechaza* H_0.

10.2. Prueba de una hipó tesis estadística

Para ilustrar los conceptos que se utilizan al probar una hipótesis estadística acerca de una población considere el siguiente ejemplo. Se sabe que, después de un periodo de dos años, cierto tipo de vacuna contra un virus que produce resfriado ya sólo es 25% eficaz. Suponga que se eligen 20 personas al azar y se les aplica una vacuna nueva, un poco más costosa, para determinar si protege contra el mismo virus durante un periodo más largo. (En un estudio real de este tipo el número de participantes que reciben la nueva vacuna podría ascender a varios miles. Aquí la muestra es de 20 sólo porque lo único que se busca es demostrar los pasos básicos para realizar una prueba estadística). Si más de 8 individuos de los que reciben la nueva vacuna superan el lapso de 2 años sin contraer el virus, la nueva vacuna se considerará superior a la que se usa en la actualidad. El requisito de que el número exceda a 8 es algo arbitrario, aunque parece razonable, ya que representa una mejoría modesta sobre las 5 personas que se esperaría recibieran protección si fueran inoculadas con la vacuna que actualmente está en uso. En esencia probamos la hipótesis nula de que la nueva vacuna es igual de eficaz después de un periodo de 2 años que la que se utiliza en la actualidad. La hipótesis alternativa es que la nueva vacuna es

mejor, y esto equivale a poner a prueba la hipótesis de que el parámetro binomial para la probabilidad de un éxito en un ensayo dado es $p = ¼$, contra la alternativa de que $p > ¼$. Esto por lo general se escribe como se indica a continuación:

$$H_0: p = 0.25,$$
$$H_1: p > 0.25.$$

El estadístico de prueba

El **estadístico de prueba** en el cual se basa nuestra decisión es X, el número de individuos en nuestro grupo de prueba que reciben protección de la nueva vacuna durante un periodo de al menos 2 años. Los valores posibles de X, de 0 a 20, se dividen en dos grupos: los números menores o iguales que 8 y aquellos mayores que 8. Todos los posibles valores mayores que 8 constituyen la **región crítica**. El último número que observamos al pasar a la región crítica se llama **valor crítico**. En nuestro ejemplo el valor crítico es el número 8. Por lo tanto, si $x > 8$, rechazamos H_0 a favor de la hipótesis alternativa H_1. Si $x \leq 8$, no rechazamos H_0. Este criterio de decisión se ilustra en la figura 10.1.

Figura 10.1: Criterio de decisión para probar $p = 0.25$ contra $p > 0.25$.

La probabilidad de un error tipo 1

El procedimiento de toma de decisiones recién descrito podría conducir a cualquiera de dos conclusiones erróneas. Por ejemplo, es probable que la nueva vacuna no sea mejor que la que se usa en la actualidad (H_0 verdadera) y, sin embargo, en este grupo específico de individuos seleccionados aleatoriamente más de 8 pasan el periodo de 2 años sin contraer el virus. Si rechazáramos H_0 a favor de H_1 cuando, de hecho, H_0 es verdadera, cometeríamos un error que se conoce como **error tipo I**.

Definición 10.2: El rechazo de la hipótesis nula cuando es verdadera se denomina **error tipo I**.

Si 8 o menos miembros del grupo superan exitosamente el periodo de 2 años y no concluimos que la nueva vacuna es mejor cuando en realidad sí lo es (H_1 verdadera), cometemos un segundo tipo de error, el de no rechazar la hipótesis H_0 cuando en realidad es falsa. A este error se le conoce como **error tipo II**.

Definición 10.3: No rechazar la hipótesis nula cuando es falsa se denomina **error tipo II**.

Al probar cualquier hipótesis estadística, hay cuatro situaciones posibles que determinan si nuestra decisión es correcta o errónea. Estas cuatro situaciones se resumen en la tabla 10.1.

Tabla 10.1: Situaciones posibles al probar una hipótesis estadística.

	H_0 es verdadera	H_0 es falsa
No rechazar H_0	Decisión correcta	Error tipo II
Rechazar H_0	Error tipo I	Decisión correcta

La probabilidad de cometer un error tipo I, también llamada **nivel de significancia**, se denota con la letra griega α. En nuestro ejemplo un error tipo I ocurriría si más de 8 individuos inoculados con la nueva vacuna superan el periodo de 2 años sin contraer el virus y los investigadores concluyen que la nueva vacuna es mejor, cuando en realidad es igual a la vacuna que se utiliza en la actualidad. Por lo tanto, si X es el número de individuos que permanecen sin contraer el virus por al menos dos años,

$$\alpha = P(\text{error tipo I}) = P\left(X > 8 \text{ cuando } p = \frac{1}{4}\right) = \sum_{x=9}^{20} b\left(x; 20, \frac{1}{4}\right)$$
$$= 1 - \sum_{x=0}^{8} b\left(x; 20, \frac{1}{4}\right) = 1 - 0.9591 = 0.0409.$$

Decimos que la hipótesis nula, $p = 1/4$, se prueba al nivel de significancia $\alpha = 0.0409$. En ocasiones el nivel de significancia se conoce como **tamaño de la prueba**. Una región crítica de tamaño 0.0409 es muy pequeña y, por lo tanto, es poco probable que se cometa un error de tipo I. En consecuencia, sería poco probable que más de 8 individuos permanecieran inmunes a un virus durante 2 años utilizando una vacuna nueva que en esencia es equivalente a la que actualmente está en el mercado.

La probabilidad de un error tipo II

La probabilidad de cometer un error tipo II, que se denota con β, es imposible de calcular a menos que tengamos una hipótesis alternativa específica. Si probamos la hipótesis nula $p = 1/4$ contra la hipótesis alternativa $p = 1/2$, entonces podremos calcular la probabilidad de no rechazar H_0 cuando es falsa. Simplemente calculamos la probabilidad de obtener 8 o menos en el grupo que supera el periodo de 2 años cuando $p = 1/2$. En este caso,

$$\beta = P\,(\text{error tipo II}) = P\left(X \leq 8 \text{ cuando } p = \frac{1}{2}\right)$$
$$= \sum_{x=0}^{8} b\left(x; 20, \frac{1}{2}\right) = 0.2517.$$

Se trata de una probabilidad elevada que indica un procedimiento de prueba en el cual es muy probable que se rechace la nueva vacuna cuando, de hecho, es mejor a la que está actualmente en uso. De manera ideal, es preferible utilizar un procedimiento de prueba con el cual haya pocas probabilidades de cometer el error tipo I y el error tipo II.

Es posible que el director del programa de prueba esté dispuesto a cometer un error tipo II si la vacuna más costosa no es significativamente mejor. De hecho, la única ocasión en la que desea evitar un error tipo II es cuando el verdadero valor de p es de al menos 0.7. Si $p = 0.7$, este procedimiento de prueba da

$$\beta = P(\text{error tipo II}) = P(X \leq 8 \text{ cuando } p = 0.7)$$
$$= \sum_{x=0}^{8} b(x; 20, 0.7) = 0.0051.$$

Con una probabilidad tan pequeña de cometer un error tipo II es muy improbable que se rechace la nueva vacuna cuando tiene una efectividad de 70% después de un periodo de 2 años. A medida que la hipótesis alternativa se aproxima a la unidad, el valor de β tiende a disminuir hasta cero.

El papel que desempeñan α, β y el tamaño de la muestra

Supongamos que el director del programa de prueba no está dispuesto a cometer un error tipo II cuando la hipótesis alternativa $p = 1/2$ es verdadera, aun cuando se encuentre que la probabilidad de tal error es $\beta = 0.2517$. Siempre es posible reducir β aumentando el tamaño de la región crítica. Por ejemplo, considere lo que les sucede a los valores de α y β cuando cambiamos nuestro valor crítico a 7, de manera que todos los valores mayores que 7 caigan en la región crítica y aquellos menores o iguales que 7 caigan en la región de no rechazo. Así, al probar $p = 1/4$ contra la hipótesis alternativa $p = 1/2$, encontramos que

$$\alpha = \sum_{x=8}^{20} b\left(x; 20, \frac{1}{4}\right) = 1 - \sum_{x=0}^{7} b\left(x; 20, \frac{1}{4}\right) = 1 - 0.8982 = 0.1018$$

$$\beta = \sum_{x=0}^{7} b\left(x; 20, \frac{1}{2}\right) = 0.1316.$$

Al adoptar un nuevo procedimiento de toma de decisiones, reducimos la probabilidad de cometer un error tipo II a costa de aumentar la probabilidad de cometer un error tipo I. Para un tamaño muestral fijo, una disminución en la probabilidad de un error por lo general tendrá como resultado un incremento en la probabilidad del otro error. Por fortuna, **la probabilidad de cometer ambos tipos de errores se puede reducir aumentando el tamaño de la muestra**. Considere el mismo problema usando una muestra aleatoria de 100 individuos. Si más de 36 miembros del grupo superan el periodo de 2 años, rechazamos la hipótesis nula de $p = 1/4$ y aceptamos la hipótesis alternativa de $p > 1/4$. El valor crítico ahora es 36. Todos los valores posibles mayores de 36 constituyen la región crítica y todos los valores posibles menores o iguales que 36 caen en la región de aceptación.

Para determinar la probabilidad de cometer un error tipo I debemos utilizar la aproximación a la curva normal con

$$\mu = np = (100)\left(\frac{1}{4}\right) = 2 \quad \text{y} \quad \sigma = \sqrt{npq} = \sqrt{(100)(1/4)(3/4)} = 4.33.$$

Con respecto a la figura 10.2, necesitamos el área bajo la curva normal a la derecha de $x = 36.5$. El valor z correspondiente es

$$z = \frac{36.5 - 25}{4.33} = 2.66.$$

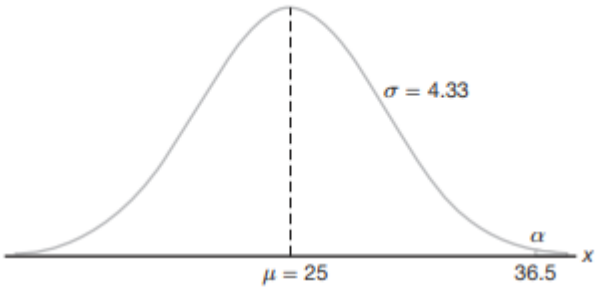


Figura 10.2: Probabilidad de un error tipo I.

En la tabla A.3 encontramos que

$$\alpha = P(\text{error tipo I}) = P\left(X > 36 \text{ cuando } p = \frac{1}{4}\right) \approx P(Z > 2.66)$$
$$= 1 - P(Z < 2.66) = 1 - 0.9961 = 0.0039.$$

Si H_0 es falsa y el verdadero valor de H_1 es $p = 1/2$, determinamos la probabilidad de un error tipo II usando la aproximación a la curva normal con

$$\mu = np = (100)(1/2) = 50 \quad \text{y} \quad \sigma = \sqrt{npq} = \sqrt{(100)(1/2)(1/2)} = 5\,.$$

La probabilidad de que un valor caiga en la región de no rechazo cuando H_0 es verdadera es dada por el área de la región sombreada a la izquierda de $x = 36.5$ en la figura 10.3. El valor z que corresponde a $x = 36.5$ es

$$z = \frac{36.5 - 50}{5} = -2.7.$$

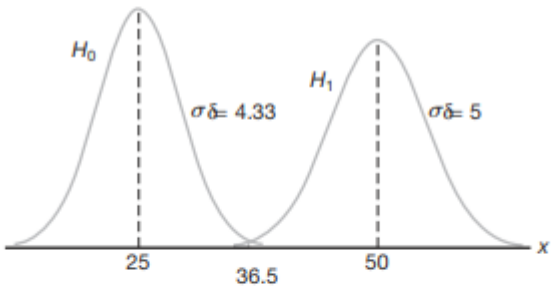


Figura 10.3: Probabilidad de un error tipo II.

Por lo tanto,

$$\beta = P(\text{error tipo II}) = P\left(X \leq 36 \text{ cuando } p = \frac{1}{2}\right) \approx P(Z < -2.7) = 0.0035.$$

Evidentemente, los errores tipo I y tipo II rara vez ocurren si el experimento consta de 100 individuos.

El ejemplo anterior destaca la estrategia del científico en la prueba de hipótesis. Después de que se plantean las hipótesis nula y alternativa es importante considerar la sensibilidad del procedimiento de prueba. Con esto queremos decir que debería determinarse un valor razonable a una α fija para la probabilidad de aceptar de manera errónea H_0, es decir, el valor de β, cuando la verdadera situación representa alguna *desviación importante de* H_0. Por lo general, es posible determinar un valor para el tamaño de la muestra, para el que existe un equilibrio razonable entre los valores de α y β que se calcula de esta manera. El problema de la vacuna es un ejemplo.

Ilustración con una variable aleatoria continua

Los conceptos que se analizan aquí para una población discreta también se pueden aplicar a variables aleatorias continuas. Considere la hipótesis nula de que el peso promedio de estudiantes hombres en cierta universidad es de 68 kilogramos, contra la hipótesis alternativa de que es diferente a 68. Es decir, deseamos probar

$$H_0: \mu = 68,$$
$$H_1: \mu \neq 68.$$

La hipótesis alternativa nos permite la posibilidad de que $\mu < 68$ o $\mu > 68$.

Una media muestral que caiga cerca del valor hipotético de 68 se consideraría como evidencia a favor de H_0. Por otro lado, una media muestral considerablemente menor que o mayor que 68 sería evidencia en contra de H_0 y, por lo tanto, favorecería a H_1. La media muestral es el estadístico de prueba en este caso. Una región crítica para el estadístico de prueba se puede elegir de manera arbitraria como los dos intervalos $\bar{x} < 67$ y $\bar{x} > 69$. La región de no rechazo será entonces el intervalo $67 \leq \bar{x} \leq 69$. Este criterio de decisión se ilustra en la figura 10.4.

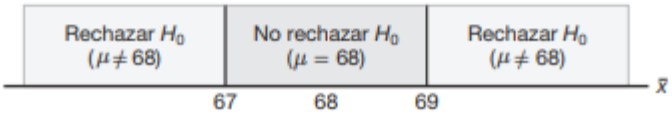


Figura 10.4: Región crítica (en azul).

Utilicemos ahora el criterio de decisión de la figura 10.4 para calcular las probabilidades de cometer los errores tipo I y tipo II cuando probemos la hipótesis nula $\mu = 68$ kilogramos contra la alternativa $\mu \neq 68$ kilogramos.

Suponga que la desviación estándar de la población de pesos es $\sigma = 3.6$. Para muestras grandes podemos sustituir s por σ si no disponemos de ninguna otra estimación de σ. Nuestro estadístico de decisión, que se basa en una muestra aleatoria de tamaño $n = 36$, será $\bar{X}$, el estimador más eficaz de μ. Del teorema del límite central sabemos que la distribución muestral de $\bar{X}$ es aproximadamente normal con desviación estándar $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 3.6/6 = 0.6$.

La probabilidad de cometer un error tipo I, o el nivel de significancia de nuestra prueba, es igual a la suma de las áreas sombreadas en cada cola de la distribución en la figura 10.5. Por lo tanto,

$$\alpha = P(\bar{X} < 67 \text{ cuando } \mu = 68) + P(\bar{X} > 69 \text{ cuando } \mu = 68).$$

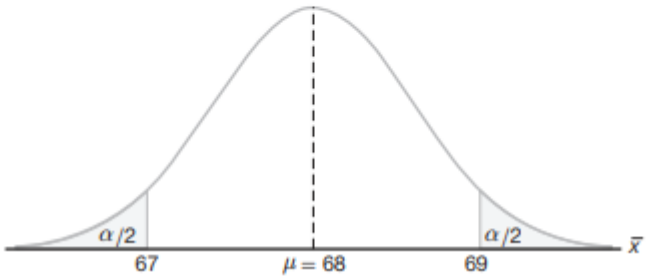


Figura 10.5: Región crítica para probar $\mu = 68$ contra $\mu \neq 68$.

Los valores z correspondientes a $\bar{x}_1 = 67$ y $\bar{x}_2 = 69$ cuando H_0 es verdadera son

$$z_1 = \frac{67 - 68}{0.6} = -1.67 \quad \text{y} \quad z_2 = \frac{69 - 68}{0.6} = 1.67.$$

Por lo tanto,

$$\alpha = P(Z < -1.67) + P(Z > 1.67) = 2P(Z < -1.67) = 0.0950.$$

Por consiguiente, 9.5% de todas las muestras de tamaño 36 nos conducirían a rechazar $\mu = 68$ kilogramos cuando, de hecho, ésta es verdadera. Para reducir α tenemos que elegir entre aumentar el tamaño de la muestra o ampliar la región de no rechazo. Suponga que aumentamos el tamaño de la muestra a $n = 64$. Entonces $\sigma_{\bar{X}} = 3.6/8 = 0.45$. En consecuencia,

$$z_1 = \frac{67-68}{0.45} = -2.22 \quad \text{y} \quad z_2 = \frac{69-68}{0.45} = 2.22.$$

Por lo tanto,

$$\alpha = P(Z < -2.22) + P(Z > 2.22) = 2\,P(Z < -2.22) = 0\,.0264.$$

La reducción de α no es suficiente por sí misma para garantizar un buen procedimiento de prueba. Debemos evaluar β para varias hipótesis alternativas. Si es importante rechazar H_0 cuando la media verdadera sea algún valor $\mu \geq 70$ o $\mu \leq 66$, entonces se debería calcular y examinar la probabilidad de cometer un error tipo II para las alternativas $\mu = 66$ y $\mu = 70$. Debido a la simetría, sólo es necesario considerar la probabilidad de no rechazar la hipótesis nula $\mu = 68$ cuando la alternativa $\mu = 70$ es verdadera. Cuando la media muestral $\bar{x}$ caiga entre 67 y 69, cuando H_1 sea verdadera, resultará un error tipo II. Por lo tanto, remitiéndonos a la figura 10.6 encontramos que

$$\beta = P(67 \leq \bar{X} \leq 69 \text{ cuando } \mu = 70).$$

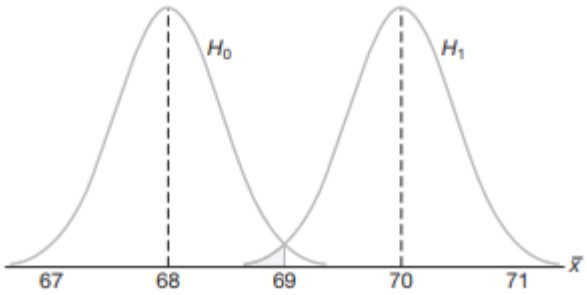


Figura 10.6: Probabilidad del error tipo II al probar $\mu = 68$ contra $\mu = 70$.

Los valores z que corresponden a $\bar{x}_1 = 67$ y $\bar{x}_2 = 69$ cuando H_1 es verdadera son

$$z_1 = \frac{67-70}{0.45} = -6.67 \quad \text{y} \quad z_2 = \frac{69-70}{0.45} = -2.22.$$

Por lo tanto,

$$\begin{aligned}\beta = P(-6.67 < Z < -2.22) &= P(Z < -2.22) - P(Z < -6.67)\\ &= 0.0132 - 0.0000 = 0.0132.\end{aligned}$$

Si el valor verdadero de μ es la alternativa $\mu = 66$, el valor de β nuevamente será 0.0132. Para todos los valores posibles de $\mu < 66$ o $\mu > 70$, el valor de β será incluso más pequeño cuando $n = 64$ y, en consecuencia, habrá poca oportunidad de no rechazar H_0 cuando sea falsa.

La probabilidad de cometer un error tipo II aumenta rápidamente cuando el valor verdadero de μ se aproxima al valor hipotético pero no es igual a éste. Desde luego, ésta suele ser la situación en la que no nos importa cometer un error tipo II. Por ejemplo, si la hipótesis alternativa $\mu = 68.5$ es verdadera, no nos importa cometer un error tipo II al concluir que la respuesta verdadera es $\mu = 68$. La probabilidad de cometer tal error será elevada cuando $n = 64$. Al remitirnos a la figura 10.7, tenemos

$$\beta = P(67 \leq \bar{X} \leq 69 \text{ cuando } \mu = 68.5).$$

Los valores z correspondientes a $\bar{x}_1 = 67$ y $\bar{x}_2 = 69$ cuando $\mu = 68.5$ son

$$z_1 = \frac{67 - 68.5}{0.45} = -3.33 \quad \text{y} \quad z_2 = \frac{69 - 68.5}{0.45} = 1.11.$$

Por lo tanto,

$$\begin{aligned}\beta = P(-3.33 < Z < 1.11) &= P(Z < 1.11) - P(Z < -3.33)\\ &= 0.8665 - 0.0004 = 0.8661.\end{aligned}$$

Los ejemplos anteriores ilustran las siguientes propiedades importantes:

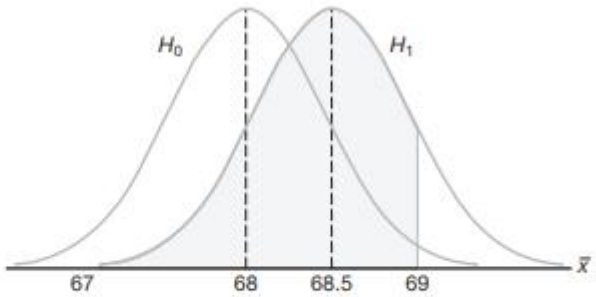


Figura 10.7: Error tipo II para la prueba de $\mu = 68$ contra $\mu = 68.5$.

Propiedades importantes de una prueba de hipótesis

1. Los errores tipo I y tipo II están relacionados. Por lo general una disminución en la probabilidad de cometer uno da como resultado un incremento en la probabilidad de cometer el otro.
2. El tamaño de la región crítica y, por lo tanto, la probabilidad de cometer un error tipo I, siempre se puede reducir ajustando el (los) valor(es) crítico(s).
3. Un aumento en el tamaño de la muestra n reducirá α y β de forma simultánea.
4. Si la hipótesis nula es falsa, β es un máximo cuando el valor verdadero de un parámetro se aproxima al valor hipotético. Cuanto más grande sea la distancia entre el valor verdadero y el valor hipotético, más pequeña será β.

Definición 10.4: La **potencia** de una prueba es la probabilidad de rechazar H_0 dado que una alternativa específica es verdadera.

10.3. Pruebas de una cola y dos colas

Una prueba de cualquier hipótesis estadística donde la alternativa es **unilateral**, como

$$H_0: \theta = \theta_0,$$
$$H_1: \theta > \theta_0,$$

o quizás

$$H_0: \theta = \theta_0,$$
$$H_1: \theta < \theta_0,$$

se denomina **prueba de una sola cola**. Anteriormente en esta sección se hizo referencia al **estadístico de prueba** para una hipótesis. Por lo general la región crítica para la hipótesis alternativa $\theta > \theta_0$ yace en la cola derecha de la distribución del estadístico de prueba, en tanto que la región crítica para la hipótesis alternativa $\theta < \theta_0$ yace por completo en la cola izquierda. (En cierto sentido el símbolo de desigualdad señala la dirección en donde se encuentra la región crítica). En el experimento de la vacuna se utilizó una prueba de una sola cola para probar la hipótesis $p = 1/4$ contra la alternativa unilateral $p > 1/4$ para la distribución binomial. La región crítica de una sola cola por lo general es evidente; el lector debería visualizar el comportamiento del estadístico de prueba y observar la *señal* evidente que produciría evidencia que respalde la hipótesis alternativa.

La prueba de cualquier hipótesis alternativa donde la alternativa es **bilateral**, como

$$H_0: \theta = \theta_0,$$
$$H_1: \theta \neq \theta_0,$$

se denomina **prueba de dos colas**, ya que la región crítica se divide en dos partes, a menudo con probabilidades iguales en cada cola de la distribución del estadístico de prueba. La hipótesis alternativa $\theta \neq \theta_0$ establece que $\theta < \theta_0$ o que $\theta > \theta_0$. Se utilizó una prueba de dos colas para probar la hipótesis nula $\mu = 68$ kilogramos contra la alternativa bilateral $\mu \neq 68$ kilogramos en el ejemplo de la población continua de los pesos de estudiantes.

¿Cómo se eligen las hipótesis nula y alternativa?

Con frecuencia la hipótesis nula H_0 se plantea usando el *signo de igualdad*. Con este método se observa claramente cómo se controla la probabilidad de cometer un error tipo I. Sin embargo, hay situaciones en que "no rechazar H_0" implica que el parámetro θ podría ser cualquier valor definido por el complemento natural de la hipótesis alternativa. Por ejemplo, en el caso de la vacuna, donde la hipótesis alternativa es H_1: $p > 1/4$, es muy posible que el no rechazo de H_0 no pueda descartar un valor de p menor que 1/4. Sin embargo, es evidente que en el caso de las pruebas de una cola la consideración más importante es el planteamiento de la alternativa.

La decisión de plantear una prueba de una cola o una de dos colas depende de la conclusión que se obtenga si se rechaza H_0. La ubicación de la región crítica sólo se puede determinar después de que se plantea H_1. Por ejemplo, al probar una medicina nueva se establece la hipótesis de que no es mejor que las medicinas similares que actualmente hay en el mercado y se prueba contra la hipótesis alternativa de que la medicina nueva es mejor. Esta hipótesis alternativa dará como resultado una prueba de una sola cola, con la región crítica en la cola derecha. Sin embargo, si deseamos comparar una nueva técnica de enseñanza con el procedimiento convencional del salón de clases, la hipótesis alternativa debe permitir que el nuevo método sea inferior o superior al procedimiento convencional. Por lo tanto, la prueba sería de dos colas con la región crítica dividida en partes iguales, de manera que caiga en los extremos de las colas izquierda y derecha de la distribución de nuestro estadístico.

Ejemplo 10.1: Un fabricante de cierta marca de cereal de arroz afirma que el contenido promedio de grasa saturada no excede a 1.5 gramos por porción. Plantee las hipótesis nula y alternativa que se utilizarán para probar esta afirmación y establezca en dónde se localiza la región crítica.

***Solución*:** La afirmación del fabricante se rechazará sólo si μ es mayor que 1.5 miligramos y no se rechazará si μ es menor o igual que 1.5 miligramos. Entonces, probamos

$$H_0: \mu = 1.5,$$
$$H_1: \mu > 1.5.$$

El hecho de no rechazar H_0 no descarta valores menores que 1.5 miligramos. Como tenemos una prueba de una cola, el símbolo mayor indica que la región crítica reside por completo en la cola derecha de la distribución de nuestro estadístico de prueba $\bar{X}$. ■

10.4. El uso de los valores P en la toma de decisiones

Al probar hipótesis en las que el estadístico de prueba es discreto, la región crítica se podría elegir de manera arbitraria y determinar su tamaño. Si α es demasiado grande, se reduce haciendo un ajuste en el valor crítico. Quizá sea necesario aumentar el tamaño de la muestra para compensar la disminución que ocurre de manera automática en la potencia de la prueba.

Por generaciones enteras de análisis estadístico se ha vuelto costumbre elegir una α de 0.05 o 0.01 y seleccionar la región crítica de acuerdo con esto. Entonces, desde luego, el rechazo o no rechazo estrictos de H_0 dependerá de esa región crítica. Por ejemplo, si la prueba es de dos colas, α se fija a un nivel de significancia de 0.05 y el estadístico de prueba implica, digamos, la distribución normal estándar, entonces se observa un valor z de los datos y la región crítica es

$$z > 1.96 \quad \text{o} \quad z < -1.96,$$

donde el valor 1.96 corresponde a $z_{0.025}$ en la tabla A.3. Un valor de z en la región crítica sugiere la aseveración: "El valor del estadístico de prueba es significativo", el cual se puede traducir al lenguaje del caso. Por ejemplo, si la hipótesis es dada por

$$H_0: \mu = 10,$$
$$H_1: \mu \neq 10,$$

Preselección de un nivel de significancia

Esta preselección de un nivel de significancia α tiene sus raíces en la filosofía de que se debe controlar el riesgo máximo de cometer un error tipo I. Sin embargo, este enfoque no explica los valores del estadístico de prueba que están "cercanos" a la región crítica. Suponga, por ejemplo, que en el caso de H_0: $\mu = 10$, contra H_1: $\mu \neq 10$, se observa un valor $z = 1.87$. En términos estrictos, con $\alpha = 0.05$ el valor no es significativo; pero el riesgo de cometer un error tipo I si se rechaza H_0 en este caso difícilmente se podría considerar grave. De hecho, en una situación de dos colas, el riesgo se cuantifica como

$$P = 2P(Z > 1.87 \text{ cuando } \mu = 10) = 2(0.0307) = 0.0614.$$

Como resultado, 0.0614 es la probabilidad de obtener un valor de z tan grande o mayor (en magnitud) que 1.87 cuando, de hecho, $\mu = 10$. Aunque esta evidencia en contra de H_0 no es tan firme como la que resultaría de un rechazo a un nivel $\alpha = 0.05$, se trata de información importante para el usuario. De hecho, el uso continuo de $\alpha = 0.05$ o 0.01 tan sólo es un resultado de lo que los estándares han transmitido por generaciones. **En la estadística aplicada los usuarios han adoptado de forma extensa el método del valor *P***. El método está diseñado para dar al usuario una alternativa (en términos de una probabilidad) a la mera conclusión de "rechazo" o "no rechazo". El cálculo del valor P también proporciona al usuario información importante cuando el valor z cae *dentro de la región crítica ordinaria.* Por ejemplo, si z es 2.73, resulta informativo para el usuario observar que

$$P = 2(0.0032) = 0.0064,$$

y, por consiguiente, el valor z es significativo a un nivel considerablemente menor que 0.05. Es importante saber que bajo la condición de H_0 un valor de $z = 2.73$ es un evento demasiado raro. A saber, un valor al menos tan grande en magnitud sólo ocurriría 64 veces en 10,000 experimentos.

Demostración gráfica de un valor *P*

Una manera muy simple de explicar gráficamente un valor P consiste en considerar dos muestras distintas. Suponga que se están considerando dos materiales para cubrir un tipo específico de metal con el fin de evitar la corrosión. Se obtienen especímenes y se cubre un grupo con el material 1 y otro grupo con el material 2. Los tamaños muestrales son $n_1 = n_2 = 10$ para cada muestra y la corrosión se mide en el porcentaje del área superficial afectada. La hipótesis plantea que las muestras provienen de distribuciones comunes con media $\mu = 10$. Supongamos que la varianza de la población es 1.0. Entonces, probamos

$$H_0 : \mu_1 = \mu_2 = 10.$$

Representemos con la figura 10.8 una gráfica de puntos de los datos. Los datos se colocan en la distribución determinada por la hipótesis nula. Supongamos que los datos "×" se refieren al material 1 y que los datos "∘" se refieren al material 2. Parece evidente que los datos realmente refutan la hipótesis nula. Pero, ¿cómo se podría resumir esto en un número? **El valor *P* se puede considerar simplemente como la probabilidad de obtener este conjunto de datos dado que las muestras provienen de la misma distribución**. Es evidente que esta probabilidad es muy pequeña, ¡digamos 0.00000001! Por consiguiente, el pequeño valor P evidentemente refuta H_0, y la conclusión es que las medias de la población son significativamente diferentes.

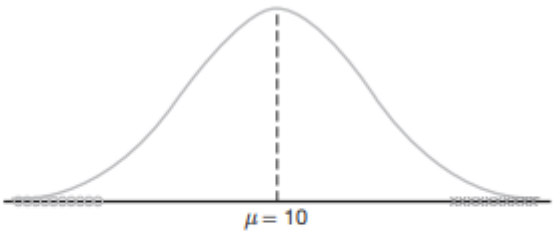


Figura 10.8: Datos que son probablemente generados de poblaciones que tienen dos medias diferentes.

El uso del método del valor P como auxiliar en la toma de decisiones es muy natural y casi todos los programas de cómputo que proporcionan el cálculo de pruebas de hipótesis ofrecen valores P, junto con valores del estadístico de prueba adecuado. La siguiente es una definición formal de un valor P.

Definición 10.5: Un **valor *P*** es el nivel (de significancia) más bajo en el que el valor observado del estadístico de prueba es significativo.

¿En qué difiere el uso de los valores *P* de la prueba de hipótesis clásica?

En este momento resulta tentador resumir los procedimientos que se asocian con la prueba de, digamos, H_0: $\theta = \theta_0$. Sin embargo, el estudiante que es novato en esta área deberá tener en cuenta que hay diferencias entre el enfoque y la filosofía del método clásico de α fija, que tiene su momento más importante en la conclusión de "rechazar H_0" o "no rechazar H_0" y el método del valor P. En este último no se determina una α fija y las conclusiones se obtienen con base en el tamaño del valor P, según la apreciación subjetiva del ingeniero o del científico. Aun cuando los modernos programas de cómputo proporcionan valores P, es importante que el lector comprenda ambos enfoques para apreciar la totalidad de los conceptos. Por lo tanto, ofrecemos una breve lista con los pasos del procedimiento tanto para el método clásico como para el del valor P.

Aproximación a la prueba de hipótesis con probabilidad fija del error tipo I	**1.** Establezca las hipótesis nula y alternativa. **2.** Elija un nivel de significancia α fijo. **3.** Seleccione un estadístico de prueba adecuado y establezca la región crítica con base en α. **4.** Rechace H_0 si el estadístico de prueba calculado está en la región crítica. De otra manera, no rechace H_0. **5.** Saque conclusiones científicas y de ingeniería.

Prueba de significancia (método del valor P)	**1.** Establezca las hipótesis nula y alternativa. **2.** Elija un estadístico de prueba adecuado. **3.** Calcule el valor P con base en los valores calculados del estadístico de prueba. **4.** Saque conclusiones con base en el valor P y los conocimientos del sistema científico.

En secciones posteriores de este capítulo y en los capítulos siguientes muchos ejemplos y ejercicios destacarán el método del valor P para obtener conclusiones científicas.

10.5. Pruebas relacionadas con una sola media (variancia conocida)

Primero deberíamos describir las suposiciones en las que se basa el experimento. El modelo para la situación subyacente se centra alrededor de un experimento con $X_1, X_2, \ldots, X_n$, que representan una muestra aleatoria de una distribución con media μ y varianza $\sigma^2 > 0$. Considere primero la hipótesis

$$H_0: \mu = \mu_0,$$
$$H_1: \mu \neq \mu_0.$$

El estadístico de prueba adecuado se debe basar en la variable aleatoria $\bar{X}$. En el capítulo 8 se presentó el teorema del límite central, el cual establece en esencia que, sin importar la distribución de X, la variable aleatoria $\bar{X}$ tiene una distribución casi normal con media μ y varianza σ^2/n para muestras de tamaño razonablemente grande. Por consiguiente, $\mu_{\bar{X}} = \mu$ y $\sigma^2_{\bar{X}} = \sigma^2/n$. Podemos determinar, entonces, una región crítica basada en el promedio muestral calculado $\bar{x}$. Ahora ya debería quedarle claro al lector que habrá una región crítica de dos colas para la prueba.

Estandarización de $\bar{X}$

Es conveniente estandarizar $\bar{X}$ e incluir de manera formal la variable aleatoria **normal estándar** Z, donde

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

Sabemos que, *bajo* H_0, es decir, si $\mu = \mu_0$, entonces $\sqrt{n}\left(\bar{X} - \mu_0\right)/\sigma$ tiene una distribución $n(x; 0, 1)$ y, por lo tanto, la expresión

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha$$

se puede utilizar para escribir una región de no rechazo adecuada. El lector debería tener en la mente que, formalmente, la región crítica se diseña para controlar α, la probabilidad de cometer un error tipo I. Debería ser evidente que se necesita una *señal de evidencia de dos colas* para apoyar H_1. Así, dado un valor calculado $\bar{x}$, la prueba formal implica rechazar H_0 si el *estadístico de prueba* z calculado cae en la región crítica que se describe a continuación.

Procedimiento de prueba para una sola media (varianza conocida)

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_{\alpha/2} \quad \text{o} \quad z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < -z_{\alpha/2}$$

Si $-z_{\alpha/2} < z < z_{\alpha/2}$, no se rechaza H_0. El rechazo de H_0, desde luego, implica la aceptación de la hipótesis alternativa $\mu \neq \mu_0$. Con esta definición de la región crítica debería quedar claro que habrá α probabilidades de rechazar H_0 (al caer en la región crítica) cuando, en realidad, $\mu = \mu_0$.

Aunque es más fácil entender la región crítica escrita en términos de z, escribimos la misma región crítica en términos del promedio calculado $\bar{x}$. Lo siguiente se puede escribir como un procedimiento de decisión idéntico:

$$\textit{rechazar } H_0 \text{ si } \bar{x} < a \text{ o } \bar{x} > b,$$

donde

$$a = \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \qquad b = \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

En consecuencia, para un nivel de significancia α, los valores críticos de la variable aleatoria z y $\bar{x}$ se presentan en la figura 10.9.

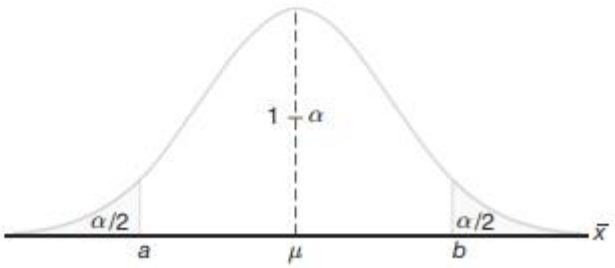


Figura 10.9: Región crítica para la hipótesis alternativa $\mu \neq \mu_0$.

Las pruebas de hipótesis unilaterales sobre la media incluyen el mismo estadístico que se describe en el caso bilateral. La diferencia, por supuesto, es que la región crítica sólo está en una cola de la distribución normal estándar. Por ejemplo, supongamos que buscamos probar

$$H_0: \mu = \mu_0,$$
$$H_1: \mu > \mu_0.$$

La señal que favorece H_1 proviene de *valores grandes* de z. Así, el rechazo de H_0 resulta cuando se calcula $z > z_\alpha$. Evidentemente, si la alternativa es $H_1: \mu < \mu_0$, la región crítica está por completo en la cola inferior, por lo que el rechazo resulta de $z < -z_\alpha$. Aunque en el caso de una prueba unilateral la hipótesis nula se puede escribir como $H_0: \mu \leq \mu_0$ o $H_0: \mu \geq \mu_0$, por lo general se escribe como $H_0: \mu = \mu_0$.

Los siguientes dos ejemplos ilustran pruebas de medias para el caso en el que se conoce σ.

10.6. Relación con la estimación del intervalo de confianza

El lector ya se habrá dado cuenta de que el método de la prueba de hipótesis para la inferencia estadística de este capítulo está muy relacionado con el método del intervalo de confianza del capítulo 9. La estimación del intervalo de confianza incluye el cálculo de límites dentro de los cuales es "razonable" que resida el parámetro en cuestión. Para el caso de una sola media de la población μ con σ^2 conocida, la estructura tanto de la prueba de hipótesis como de la estimación del intervalo de confianza se basa en la variable aleatoria

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

Resulta que la prueba de H_0: $\mu = \mu_0$ contra H_1: $\mu \neq \mu_0$ a un nivel de significancia α es *equivalente a calcular un intervalo de confianza del* $100(1-\alpha)\%$ *sobre* μ *y rechazar* H_0, *si* μ_0 *está fuera del intervalo de confianza.* Si μ_0 está dentro del intervalo de confianza, no se rechaza la hipótesis. La equivalencia es muy intuitiva y se puede ilustrar de manera muy simple. Recuerde que con un valor observado $\bar{x}$, no rechazar H_0 a un nivel de significancia α implica que

$$-z_{\alpha/2} \leq \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \leq z_{\alpha/2},$$

que es equivalente a

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

La equivalencia de la estimación del intervalo de confianza con la prueba de hipótesis se extiende a las diferencias entre dos medias, varianzas, cocientes de varianzas, etcétera. Como resultado, el estudiante de estadística no debería considerar la estimación del intervalo de confianza y la prueba de hipótesis como formas separadas de inferencia estadística. Considere el ejemplo 9.2 de la página 271. El intervalo de confianza del 95% sobre la media es dado por los límites (2.50, 2.70). Por consiguiente, con la misma información muestral, no se rechazará una hipótesis bilateral sobre μ que incluya cualquier valor hipotético entre 2.50 y 2.70. A medida que exploremos diferentes áreas de la prueba de hipótesis seguiremos aplicando la equivalencia a la estimación del intervalo de confianza.

10.1. Pruebas relacionadas con una sola media (variancia desconocida)

Ciertamente sospecharíamos que las pruebas sobre una media de la población μ con σ^2 desconocida, como la estimación del intervalo de confianza, deberían incluir el uso de la distribución t de Student. En términos estrictos, la aplicación de la t de Student tanto para los intervalos de confianza como para la prueba de hipótesis se desarrolla bajo los siguientes supuestos. Las variables aleatorias $X_1, X_2, \ldots, X_n$ representan una muestra aleatoria de una distribución normal con μ y σ^2 desconocidas. Entonces, la variable aleatoria $\sqrt{n}(\bar{X}-\mu)/S$ tiene una distribución t de Student con $n-1$ grados de libertad. La estructura de la prueba es idéntica a la del caso en el que se conoce σ, excepto que el valor σ en el estadístico de prueba se reemplaza con el estimado calculado de S y la distribución normal estándar se reemplaza con una distribución t.

El estadístico t para una prueba sobre una sola media (varianza desconocida)

Para la hipótesis bilateral

$$H_0: \mu = \mu_0,$$
$$H_1: \mu \neq \mu_0,$$

rechazamos H_0 a un nivel de significancia α cuando el estadístico t calculado

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

excede a $t_{\alpha/2,n-1}$ o es menor que $-t_{\alpha/2,n-1}$.

El lector debería recordar de los capítulos 8 y 9 que la distribución t es simétrica alrededor del valor cero. Así, esta región crítica de dos colas se aplica de manera similar a la del caso en que se conoce σ. Para la hipótesis bilateral a un nivel de significancia α se aplican las regiones críticas de dos colas. Para $H_1: \mu > \mu_0$ el rechazo resulta cuando $t > t_{\alpha,n-1}$. Para $H_1: \mu < \mu_0$ la región crítica es dada por $T < -t_{\alpha,n-1}$.

Ejemplo 10.5: El Edison Electric Institute publica cifras del número de kilowatts-hora que gastan anualmente varios aparatos electrodomésticos. Se afirma que una aspiradora gasta un promedio de 46 kilowatts-hora al año. Si una muestra aleatoria de 12 hogares, que se incluye en un estudio planeado, indica que las aspiradoras gastan un promedio de 42 kilowatts-hora al año con una desviación estándar de 11.9 kilowatts-hora, ¿esto sugiere que las aspiradoras gastan, en promedio, menos de 46 kilowatts-hora al año a un nivel de significancia de 0.05? Suponga que la población de kilowatts-hora es normal.

Solución:

1. $H_0: \mu = 46$ kilowatts-hora.
2. $H_1: \mu < 46$ kilowatts-hora.
3. $\alpha = 0.05$.
4. Región crítica: $t < -1.796$, donde $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ con 11 grados de libertad.
5. Cálculos: $\bar{x} = 42$ kilowatts-hora, $s = 11.9$ kilowatts-hora y $n = 12$.
   En consecuencia,

$$t = \frac{42 - 46}{11.9/\sqrt{12}} = -1.16, \qquad P = P(T < -1.16) \approx 0.135.$$

6. Decisión: no rechazar H_0 y concluir que el número promedio de kilowatts-hora que gastan al año las aspiradoras domésticas no es significativamente menor que 46. ∎

10.2. Pruebas sobre dos medias (Dos muestras)

El lector deberá comprender la relación entre pruebas e intervalos de confianza y sólo puede confiar plenamente en los detalles que ofrece el material sobre el intervalo de confianza del capítulo 9. Las pruebas respecto a dos medias representan un conjunto de herramientas analíticas muy importantes para el científico o el ingeniero. El procedimiento experimental es muy parecido al que se describe en la sección 9.8. Se extraen dos muestras aleatorias independientes de tamaños n_1 y n_2, respectivamente, de dos poblaciones con medias μ_1 y μ_2, y varianzas σ_1^2 y σ_2^2. Sabemos que la variable aleatoria

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

tiene una distribución normal estándar. Suponemos aquí que n_1 y n_2 son suficientemente grandes, por lo que se aplica el teorema del límite central. Por supuesto, si las dos poblaciones son normales, el estadístico anterior tiene una distribución normal estándar incluso para n_1 y n_2 pequeñas. Evidentemente, si podemos suponer que $\sigma_1 = \sigma_2 = \sigma$, el estadístico anterior se reduce a

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n_1 + 1/n_2}}.$$

Los dos estadísticos anteriores sirven como base para el desarrollo de los procedimientos de prueba que incluyen dos medias. La equivalencia entre las pruebas y los intervalos de confianza, junto con los detalles técnicos implicados en las pruebas sobre una media, permiten que la transición a pruebas con dos medias sea sencilla.

La hipótesis bilateral sobre dos medias se escribe de manera muy general como

$$H_0: \mu_1 - \mu_2 = d_0.$$

Es evidente que la alternativa puede ser bilateral o unilateral. De nuevo, la distribución que se utiliza es la distribución del estadístico de prueba bajo H_0. Se calculan los valores $\bar{x}_1$ y $\bar{x}_2$, y para σ_1 y σ_2 conocidas, el estadístico de prueba es dado por

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}},$$

con una región crítica de dos colas en el caso de una alternativa bilateral. Es decir, se rechaza H_0 a favor de H_1: $\mu_1 - \mu_2 \neq d_0$, si $z > z_{\alpha/2}$ o $z < -z_{\alpha/2}$. Las regiones críticas de una cola se utilizan en el caso de alternativas unilaterales. El lector debería estudiar, como antes, el estadístico de prueba y estar satisfecho de que para, digamos H_1: $\mu_1 - \mu_2 > d_0$, la señal que favorece H_1 provenga de valores grandes de z. Por consiguiente, se aplica la región crítica de la cola superior.

Varianzas desconocidas pero iguales

Las situaciones más comunes que implican pruebas sobre dos medias son aquellas con varianzas desconocidas. Si el científico interesado está dispuesto a suponer que ambas distribuciones son normales y que $\sigma_1 = \sigma_2 = \sigma$, se puede utilizar la *prueba t agrupada* (a menudo llamada prueba t de dos muestras). El estadístico de prueba (véase la sección 9.8) es dado por el siguiente procedimiento de prueba.

Prueba t agrupada de dos muestras

Para la hipótesis bilateral

$$H_0: \mu_1 = \mu_2,$$
$$H_1: \mu_1 \neq \mu_2,$$

rechazamos H_0 al nivel de significancia α cuando el estadístico t calculado

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p\sqrt{1/n_1 + 1/n_2}},$$

donde

$$s_p^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}$$

excede a $t_{\alpha/2,\, n_1+n_2-2}$ o es menor que $-t_{\alpha/2,\, n_1+n_2-2}$.

Recuerde que en el capítulo 9 se explicó que los grados de libertad para la distribución t son un resultado del agrupamiento de la información de las dos muestras para estimar σ^2. Las alternativas unilaterales, como era de esperarse, sugieren regiones críticas unilaterales. Por ejemplo, para H_1: $\mu_1 - \mu_2 > d_0$, rechace H_1: $\mu_1 - \mu_2 = d_0$ cuando $t > t_{\alpha, n_1+n_2-2}$.

Varianzas desconocidas pero diferentes

Hay situaciones donde al analista **no** le es posible suponer que $\sigma_1 = \sigma_2$. De la sección 9.8 recuerde que, si las poblaciones son normales, el estadístico

$$T' = \frac{(\bar{X}_1 - \bar{X}_2) - d_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

tiene una distribución t aproximada con grados de libertad aproximados

$$v = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}.$$

Como resultado, el procedimiento de prueba consiste en *no rechazar* H_0 cuando

$$-t_{\alpha/2,v} < t' < t_{\alpha/2,v},$$

con v dado como antes. De nuevo, como en el caso de la prueba t agrupada, las alternativas unilaterales sugieren regiones críticas unilaterales.

10.3. Selección del tamaño de la muestra para la prueba de medias

En la sección 10.2 demostramos cómo el analista puede explotar las relaciones entre el tamaño de la muestra, el nivel de significancia α y la potencia de la prueba para alcanzar cierto estándar de calidad. En la mayoría de las circunstancias prácticas el experimento debería planearse y, de ser posible, elegir el tamaño de la muestra antes del proceso de recolección de datos. Por lo general el tamaño de la muestra se determina de modo que

Tabla 10.3: Pruebas relacionadas con medias

H_0	Valor del estadístico de prueba	H_1	Región crítica
$\mu = \mu_0$	$z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$; σ conocida	$\mu < \mu_0$ $\mu > \mu_0$ $\mu \neq \mu_0$	$z < -z_\alpha$ $z > z_\alpha$ $z < -z_{\alpha/2}$ o $z > z_{\alpha/2}$
$\mu = \mu_0$	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$; $v = n - 1$, σ desconocida	$\mu < \mu_0$ $\mu > \mu_0$ $\mu \neq \mu_0$	$t < -t_\alpha$ $t > t_\alpha$ $t < -t_{\alpha/2}$ o $t > t_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$z = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$; σ_1 y σ_2 conocidas	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$	$z < -z_\alpha$ $z > z_\alpha$ $z < -z_{\alpha/2}$ o $z > z_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$t = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p\sqrt{1/n_1 + 1/n_2}}$; $v = n_1 + n_2 - 2$, $\sigma_1 = \sigma_2$ pero desconocidas $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$	$t < -t_\alpha$ $t > t_\alpha$ $t < -t_{\alpha/2}$ o $t > t_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$t' = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$; $v = \dfrac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$, $\sigma_1 \neq \sigma_2$ y desconocidas	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$	$t' < -t_\alpha$ $t' > t_\alpha$ $t' < -t_{\alpha/2}$ o $t' > t_{\alpha/2}$
$\mu_D = d_0$ observaciones pareadas	$t = \dfrac{\bar{d} - d_0}{s_d/\sqrt{n}}$; $v = n - 1$	$\mu_D < d_0$ $\mu_D > d_0$ $\mu_D \neq d_0$	$t < -t_\alpha$ $t > t_\alpha$ $t < -t_{\alpha/2}$ o $t > t_{\alpha/2}$

permita lograr una buena potencia para una α fija y una alternativa específica fija. Esta alternativa fija puede estar en la forma de $\mu - \mu_0$ en el caso de una hipótesis que incluya una sola media o $\mu_1 - \mu_2$ en el caso de un problema que implique dos medias. Los casos específicos serán ilustrativos.

Suponga que deseamos probar la hipótesis

$$H_0: \mu = \mu_0,$$
$$H_1 = \mu > \mu_0,$$

con un nivel de significancia α, cuando se conoce la varianza σ^2. Para una alternativa específica, digamos, $\mu = \mu_0 + \delta$, en la figura 10.14 se muestra que la potencia de nuestra prueba es

$$1 - \beta = P(\bar{X} > a \text{ cuando } \mu = \mu_0 + \delta).$$

Por lo tanto,

$$\begin{aligned}\beta &= P(\bar{X} < a \text{ cuando } \mu = \mu_0 + \delta)\\ &= P\left[\frac{\bar{X} - (\mu_0 + \delta)}{\sigma/\sqrt{n}} < \frac{a - (\mu_0 + \delta)}{\sigma/\sqrt{n}} \text{ cuando } \mu = \mu_0 + \delta\right].\end{aligned}$$

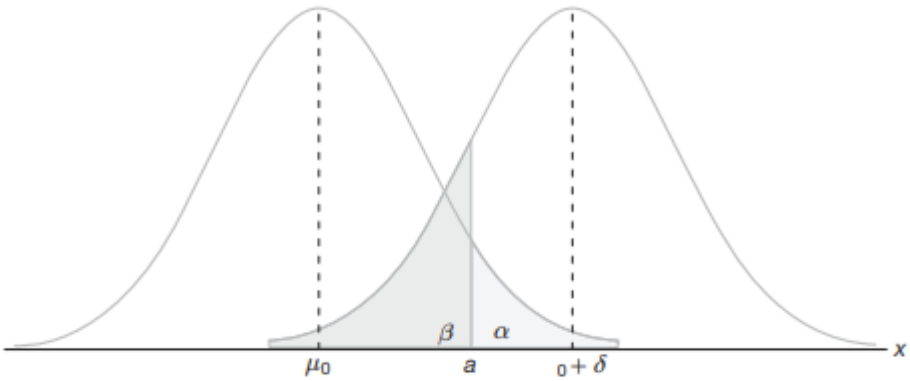


Figura 10.14: Prueba de $\mu = \mu_0$ contra $\mu = \mu_0 + \delta$.

Bajo la hipótesis alternativa $\mu = \mu_0 + \delta$, el estadístico

$$\frac{\bar{X} - (\mu_0 + \delta)}{\sigma/\sqrt{n}}$$

es la variable normal estándar Z. Por lo tanto,

$$\beta = P\left(Z < \frac{a - \mu_0}{\sigma/\sqrt{n}} - \frac{\delta}{\sigma/\sqrt{n}}\right) = P\left(Z < z_\alpha - \frac{\delta}{\sigma/\sqrt{n}}\right),$$

de donde concluimos que

$$-z_\beta = z_\alpha - \frac{\delta\sqrt{n}}{\sigma},$$

y, en consecuencia,

$$\text{Elección del tamaño de la muestra: } n = \frac{(z_\alpha + z_\beta)^2\sigma^2}{\delta^2},$$

un resultado que también es verdadero cuando la hipótesis alternativa es $\mu < \mu_0$.

En el caso de una prueba de dos colas obtenemos la potencia $1 - \beta$ para una alternativa específica cuando

$$n \approx \frac{(z_{\alpha/2} + z_\beta)^2\sigma^2}{\delta^2}.$$

El caso de dos muestras

Se puede utilizar un procedimiento similar para determinar el tamaño de la muestra $n = n_1 = n_2$ que se requiere para una potencia específica de la prueba en que se comparan dos medias de la población. Por ejemplo, suponga que deseamos probar la hipótesis

$$H_0: \mu_1 - \mu_2 = d_0,$$
$$H_1: \mu_1 - \mu_2 \neq d_0,$$

cuando se conocen σ_1 y σ_2. Para una alternativa específica, digamos $\mu_1 - \mu_2 = d_0 + \delta$, en la figura 10.15 se muestra que la potencia de nuestra prueba es

$$1 - \beta = P(|\bar{X}_1 - \bar{X}_2| > a \text{ cuando } \mu_1 - \mu_2 = d_0 + \delta).$$

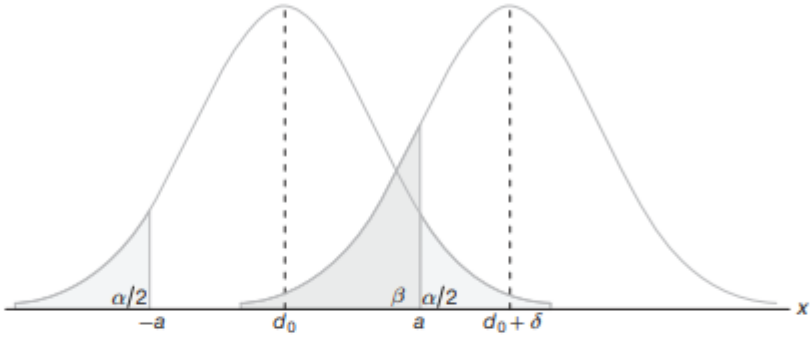


Figura 10.15: Prueba de $\mu_1 - \mu_2 = d_0$ contra $\mu_1 - \mu_2 = d_0 + \delta$.

Por lo tanto,

$$\beta = P(-a < \bar{X}_1 - \bar{X}_2 < a \text{ cuando } \mu_1 - \mu_2 = d_0 + \delta)$$
$$= P\left[\frac{-a - (d_0 + \delta)}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}} < \frac{(\bar{X}_1 - \bar{X}_2) - (d_0 + \delta)}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}}\right.$$

Por lo tanto,

$$\beta = P(-a < \bar{X}_1 - \bar{X}_2 < a \text{ cuando } \mu_1 - \mu_2 = d_0 + \delta)$$
$$= P\left[\frac{-a - (d_0 + \delta)}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}} < \frac{(\bar{X}_1 - \bar{X}_2) - (d_0 + \delta)}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}}\right.$$

$$\left. < \frac{a - (d_0 + \delta)}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}} \text{ cuando } \mu_1 - \mu_2 = d_0 + \delta\right].$$

Con la hipótesis alternativa $\mu_1 - \mu_2 = d_0 + \delta$, el estadístico

$$\frac{\bar{X}_1 - \bar{X}_2 - (d_0 + \delta)}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}}$$

es la variable normal estándar Z. Ahora bien, al escribir

$$-z_{\alpha/2} = \frac{-a - d_0}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}} \quad \text{y} \quad z_{\alpha/2} = \frac{a - d_0}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}},$$

tenemos

$$\beta = P\left[-z_{\alpha/2} - \frac{\delta}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}} < Z < z_{\alpha/2} - \frac{\delta}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}}\right],$$

de donde concluimos que

$$-z_\beta \approx z_{\alpha/2} - \frac{\delta}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}},$$

y, por lo tanto,

$$n \approx \frac{(z_{\alpha/2} + z_\beta)^2(\sigma_1^2 + \sigma_2^2)}{\delta^2}.$$

Para la prueba de una sola cola, la expresión para el tamaño requerido de la muestra cuando $n = n_1 = n_2$ es

$$\text{Elección del tamaño de la muestra: } n = \frac{(z_\alpha + z_\beta)^2(\sigma_1^2 + \sigma_2^2)}{\delta^2}.$$

Cuando se desconoce la varianza de la población (o varianzas en la situación de dos muestras), la elección del tamaño de la muestra no es directa. Al probar la hipótesis $\mu = \mu_0$ cuando el valor verdadero es $\mu = \mu_0 + \delta$, el estadístico

$$\frac{\bar{X} - (\mu_0 + \delta)}{S/\sqrt{n}}$$

no sigue la distribución t, como se podría esperar, más bien sigue la **distribución t no central**. Sin embargo, existen tablas o gráficas que se basan en la distribución t no central para determinar el tamaño adecuado de la muestra, si se dispone de algún estimado de σ o si δ es un múltiplo de σ. La tabla A.8 proporciona los tamaños muestrales necesarios para controlar los valores de α y β para diversos valores de

$$\Delta = \frac{|\delta|}{\sigma} = \frac{|\mu - \mu_0|}{\sigma}$$

en el caso de pruebas de una y de dos colas. En el caso de la prueba t de dos muestras en la que se desconocen las varianzas pero se suponen iguales, obtenemos los tamaños muestrales $n = n_1 = n_2$ necesarios para controlar los valores de α y β para diversos valores de

$$\Delta = \frac{|\delta|}{\sigma} = \frac{|\mu_1 - \mu_2 - d_0|}{\sigma}$$

de la tabla A.9.

10.4. Pruebas relacionadas con proporciones

Las pruebas de hipótesis que se relacionan con proporciones se requieren en muchas áreas. A los políticos les interesa conocer la fracción de votantes que los favorecerá en la siguiente elección. Todas las empresas manufactureras se preocupan por la proporción de artículos defectuosos cuando se realiza un embarque. Los jugadores dependen del conocimiento de la proporción de resultados que consideran favorables.

Consideraremos el problema de probar la hipótesis de que la proporción de éxitos en un experimento binomial es igual a algún valor específico. Es decir, probaremos la hipótesis nula H_0 de que $p = p_0$, donde p es el parámetro de la distribución binomial. La hipótesis alternativa puede ser una de las alternativas unilaterales o bilaterales usuales:

$$p < p_0, \quad p > p_0, \quad \text{o} \quad p \neq p_0.$$

La variable aleatoria adecuada sobre la que basamos nuestro criterio de decisión es la variable aleatoria binomial X; aunque también podríamos usar el estadístico $\hat{p} = X/n$. Los valores de X que están lejos de la media $\mu = np_0$ conducirán al rechazo de la hipótesis nula. Como X es una variable binomial discreta, es poco probable que se pueda establecer una región crítica cuyo tamaño sea *exactamente* igual a un valor preestablecido de α. Por esta razón es preferible, al trabajar con muestras pequeñas, basar nuestras decisiones en valores P. Para probar la hipótesis

$$H_0: p = p_0,$$
$$H_1: p < p_0,$$

utilizamos la distribución binomial para calcular el valor P

$$P = P(X \leq x \text{ cuando } p = p_0).$$

El valor x es el número de éxitos en nuestra muestra de tamaño n. Si este valor P es menor o igual que α, nuestra prueba es significativa al nivel α y rechazamos H_0 a favor de H_1. De manera similar, para probar la hipótesis

$$H_0: p = p_0,$$
$$H_1: p > p_0,$$

al nivel de significancia α, calculamos

$$P = P(X \geq x \text{ cuando } p = p_0)$$

y rechazamos H_0 a favor de H_1 si este valor P es menor o igual que α. Finalmente, para probar la hipótesis
a un nivel de significancia α, calculamos

$$P = 2P(X \leq x \text{ cuando } p = p_0) \qquad \text{si } x < np_0$$

o

$$P = 2P(X \geq x \text{ cuando } p = p_0) \qquad \text{si } x > np_0$$

y rechazamos H_0 a favor de H_1 si el valor P calculado es menor o igual que α.

Los pasos para probar una hipótesis nula acerca de una proporción contra varias alternativas usando las probabilidades binomiales de la tabla A.1 son los siguientes:

Prueba de una proporción (muestras pequeñas)

1. $H_0: p = p_0$.
2. Una de las alternativas $H_1: p < p_0$, $p > p_0$ o $p \neq p_0$.
3. Elegir un nivel de significancia igual a α.
4. Estadístico de prueba: variable binomial X con $p = p_0$.
5. Cálculos: obtener x, el número de éxitos, y calcular el valor P adecuado.
6. Decisión: sacar las conclusiones apropiadas con base en el valor P.

Dos muestras: pruebas sobre dos proporciones

A menudo surgen situaciones en las que se desea probar la hipótesis de que dos proporciones son iguales. Por ejemplo, podemos tratar de mostrar evidencia de que la proporción de médicos que son pediatras en un estado es igual a la proporción de pediatras en otro estado. Quizás un individuo decida dejar de fumar sólo si se convence de que la proporción de fumadores con cáncer pulmonar excede a la proporción de no fumadores con ese tipo de cáncer.

En general, deseamos probar la hipótesis nula de que dos proporciones, o parámetros binomiales, son iguales. Es decir, probamos $p_1 = p_2$ contra una de las alternativas $p_1 < p_2$, $p_1 > p_2$, o $p_1 \neq p_2$. Desde luego, esto es equivalente a probar la hipótesis nula de que $p_1 - p_2 = 0$ contra una de las alternativas $p_1 - p_2 < 0$, $p_1 - p_2 > 0$ o $p_1 - p_2 \neq 0$. El estadístico sobre el que basamos nuestra decisión es la variable aleatoria $\hat{P}_1 - \hat{P}_2$. Se seleccionan al azar muestras independientes de tamaños n_1 y n_2 de dos poblaciones binomiales y se calcula la proporción de éxitos $\hat{P}_1$ y $\hat{P}_2$ para las dos muestras.

En la construcción de intervalos de confianza para p_1 y p_2 observamos, para n_1 y n_2 suficientemente grandes, que el estimador puntual $\hat{P}_1$ menos $\hat{P}_2$ estaba distribuido de forma casi normal con media

$$\mu_{\hat{P}_1 - \hat{P}_2} = p_1 - p_2$$

y varianza

$$\sigma^2_{\hat{P}_1 - \hat{P}_2} = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}.$$

Por lo tanto, es posible establecer la(s) región(es) crítica(s) usando la variable normal estándar

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\sqrt{p_1 q_1 / n_1 + p_2 q_2 / n_2}}.$$

Cuando H_0 es verdadera, podemos sustituir $p_1 = p_2 = p$ y $q_1 = q_2 = q$ (donde p y q son los valores comunes) en la fórmula anterior para Z y obtener la forma

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{pq(1/n_1 + 1/n_2)}}.$$

Sin embargo, para calcular un valor de Z debemos estimar los parámetros p y q que aparecen en el radical. Al agrupar los datos de ambas muestras el **estimado agrupado de la proporción** p es

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2},$$

donde x_1 y x_2 son el número de éxitos en cada una de las dos muestras. Al sustituir $\hat{p}$ por p y $\hat{q} = 1 - \hat{p}$ por q, el **valor z para probar $p_1 = p_2$** se determina a partir de la fórmula

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}(1/n_1 + 1/n_2)}}.$$

Las regiones críticas para las hipótesis alternativas adecuadas se establecen como antes, utilizando puntos críticos de la curva normal estándar. En consecuencia, para la alternativa $p_1 \neq p_2$, al nivel de significancia α, la región crítica es $z < -z_{\alpha/2}$ o $z > z_{\alpha/2}$. Para una prueba donde la alternativa es $p_1 < p_2$, la región crítica será $z < -z_\alpha$; y cuando la alternativa es $p_1 > p_2$, la región crítica será $z > z_\alpha$.

10.5. Pruebas de una y dos muestras referentes a varianzas

En esta sección estudiaremos la prueba de hipótesis relacionada con varianzas o desviaciones estándar de la población. No son poco comunes las aplicaciones de pruebas de una y dos muestras sobre varianzas. Los ingenieros y los científicos constantemente se enfrentan a estudios donde se les pide demostrar que las mediciones que tienen que ver con productos o procesos cumplen con las especificaciones que fijan los consumidores. Las especificaciones a menudo se cumplen si la varianza del proceso es suficientemente pequeña. También existe interés por experimentos que comparan métodos o procesos donde la reproducibilidad o variabilidad inherentes se deben comparar de manera formal. Además, para determinar si no se cumple la suposición de varianzas iguales, con frecuencia se aplica una prueba que compara dos varianzas antes de llevar a cabo una prueba t sobre dos medias.

Empecemos por considerar el problema de probar la hipótesis nula H_0 de que la varianza de la población σ^2 es igual a un valor específico σ_0^2 contra una de las alternativas comunes $\sigma^2 < \sigma_0^2$, $\sigma^2 > \sigma_0^2$ o $\sigma^2 \neq \sigma_0^2$. El estadístico apropiado sobre el que basamos nuestra decisión es el estadístico chi cuadrada del teorema 8.4, el cual se utilizó en el capítulo 9 para construir un intervalo de confianza para σ^2. Por lo tanto, si suponemos que la distribución de la población que se muestrea es normal, el valor de chi cuadrada para probar $\sigma^2 = \sigma_0^2$ es dado por

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2},$$

donde n es el tamaño de la muestra, s^2 es la varianza muestral y σ_0^2 es el valor de σ^2 dado por la hipótesis nula. Si H_0 es verdadera, χ^2 es un valor de la distribución chi cuadrada con $v = n - 1$ grados de libertad. En consecuencia, para una prueba de dos colas a un nivel de significancia α, la región crítica es $\chi^2 < \chi^2_{1-\alpha/2}$ o $\chi^2 > \chi^2_{\alpha/2}$. Para la alternativa unilateral $\sigma^2 < \sigma_0^2$, la región crítica es $\chi^2 < \chi^2_{1-\alpha}$; y para la alternativa unilateral $\sigma^2 > \sigma_0^2$, la región crítica es $\chi^2 > \chi^2_{\alpha}$.

Robustez de la prueba χ^2 para la suposición de normalidad

Tal vez el lector se habrá dado cuenta de que varias pruebas dependen, al menos en teoría, de la suposición de normalidad. En general muchos procedimientos en estadística aplicada tienen fundamentos teóricos que dependen de la distribución normal. Estos procedimientos varían en el grado en que dependen de la suposición de la normalidad. A un procedimiento que es razonablemente insensible a esta suposición se le denomina **procedimiento robusto**, es decir, robusto para la normalidad. La prueba χ^2 sobre una sola varianza no es robusta en absoluto para la normalidad, es decir, el éxito práctico del procedimiento depende de la normalidad. Como resultado, el valor P calculado podría ser notoriamente diferente del valor P verdadero si la población de la que se toma la muestra no es normal. De hecho, resulta muy plausible que un valor P estadísticamente significativo no sea una verdadera señal de H_1: $\sigma \neq \sigma_0$, sino que un valor significativo sea el resultado de haber violado las suposiciones de normalidad. Por lo tanto, el analista debería utilizar esta prueba χ^2 específica con precaución.

Ejemplo 10.12: Un fabricante de baterías para automóvil afirma que la duración de sus baterías se distribuye de forma aproximadamente normal con una desviación estándar igual a 0.9 años. Si una muestra aleatoria de 10 de tales baterías tiene una desviación estándar de 1.2 años, ¿considera que $\sigma > 0.9$ años? Utilice un nivel de significancia de 0.05.

Solución:

1. H_0: $\sigma^2 = 0.81$.
2. H_1: $\sigma^2 > 0.81$.
3. $\alpha = 0.05$.
4. Región crítica: En la figura 10.19 vemos que se rechaza la hipótesis nula cuando $\chi^2 > 16.919$, donde $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$ con $v = 9$ grados de libertad.

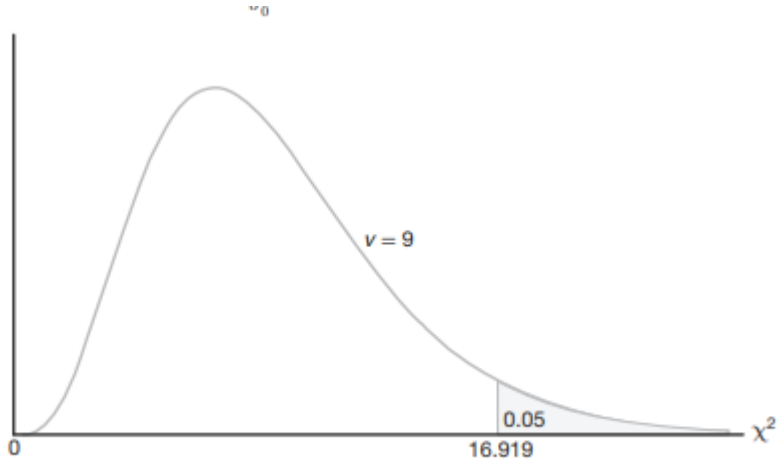


Figura 10.19: Región crítica para la hipótesis alternativa $\sigma > 0.9$.

5. Cálculos: $s^2 = 1.44$, $n = 10$ y

$$\chi^2 = \frac{(9)(1.44)}{0.81} = 16.0, \qquad P \approx 0.07.$$

6. Decisión: El estadístico χ^2 no es significativo al nivel 0.05. Sin embargo, con base en el valor P de 0.07, hay evidencia de que $\sigma > 0.9$. ∎

10.6. Prueba de la bondad de ajuste

A lo largo de este capítulo nos ocupamos de la prueba de hipótesis estadística acerca de parámetros de una sola población, como μ, σ^2 y p. Ahora consideraremos una prueba para determinar si una población tiene una distribución teórica específica. La prueba se basa en el nivel de ajuste que existe entre la frecuencia de ocurrencia de las observaciones en una muestra observada y las frecuencias esperadas que se obtienen a partir de la distribución hipotética.

Para ilustrar lo anterior considere el lanzamiento de un dado. Suponemos que se trata de un dado legal, lo cual equivale a probar la hipótesis de que la distribución de resultados es la distribución uniforme discreta

$$f(x) = \frac{1}{6}, \qquad x = 1, 2, \ldots, 6.$$

Suponga que el dado se lanza 120 veces y que se registra cada resultado. Teóricamente, si el dado está balanceado, esperaríamos que cada cara ocurriera 20 veces. Los resultados se presentan en la tabla 10.4.

Tabla 10.4: Frecuencias observadas y esperadas de 120 lanzamientos de un dado

Cara	**1**	**2**	**3**	**4**	**5**	**6**
Observadas	20	22	17	18	19	24
Esperadas	20	20	20	20	20	20

Al comparar las frecuencias observadas con las frecuencias esperadas correspondientes debemos decidir si es posible que tales discrepancias ocurran como resultado de fluctuaciones del muestreo, de que el dado está balanceado o no es legal o de que la distribución de resultados no es uniforme. Es práctica común referirse a cada resultado posible de un experimento como una celda. En nuestro caso tenemos 6 celdas. A continuación se define el estadístico adecuado en el cual basamos nuestro criterio de decisión para un experimento que incluye k celdas.

Una **prueba de la bondad de ajuste** entre las frecuencias observadas y esperadas se basa en la cantidad.

Prueba de la bondad de ajuste

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i},$$

donde χ^2 es un valor de una variable aleatoria cuya distribución muestral se aproxima muy de cerca a la distribución chi cuadrada con $v = k - 1$ grados de libertad. Los símbolos o_i y e_i representan las frecuencias observada y esperada, respectivamente, para la i-ésima celda.

El número de grados de libertad asociado con la distribución chi cuadrada que se utiliza aquí es igual a $k - 1$, pues sólo hay $k - 1$ frecuencias de celdas libremente determinadas. Es decir, una vez que se determinan las frecuencias de $k - 1$ celdas, también se determina la frecuencia para la k-ésima celda.

Si las frecuencias observadas se acercan a las frecuencias esperadas correspondientes, el valor χ^2 será pequeño, lo cual indica un buen ajuste. Si las frecuencias observadas difieren de manera considerable de las frecuencias esperadas, el valor χ^2 será grande y el ajuste deficiente. Un buen ajuste conduce a la aceptación de H_0, mientras que un ajuste deficiente conduce a su rechazo. Por lo tanto, la región crítica caerá en la cola derecha de la distribución chi cuadrada. Para un nivel de significancia igual a α encontramos el valor crítico χ^2_α de la tabla A.5 y, entonces, $\chi^2 > \chi^2_\alpha$ constituye la región crítica. **El criterio de decisión que aquí se describe no se debería utilizar a menos que cada una de las frecuencias esperadas sea por lo menos igual a 5**. Esta restricción podría requerir la combinación de celdas adyacentes, lo que dará como resultado una reducción en el número de grados de libertad.

En la tabla 10.4 encontramos que el valor χ^2 es

$$\chi^2 = \frac{(20-20)^2}{20} + \frac{(22-20)^2}{20} + \frac{(17-20)^2}{20} + \frac{(18-20)^2}{20} + \frac{(19-20)^2}{20} + \frac{(24-20)^2}{20} = 1.7.$$

Si usamos la tabla A.5, encontramos $\chi^2_{0.05} = 11.070$ para $v = 5$ grados de libertad. Como 1.7 es menor que el valor crítico, no se rechaza H_0. Concluimos que no hay suficiente evidencia de que el dado está desbalanceado.

Como un segundo ejemplo probemos la hipótesis de que la distribución de frecuencias de la duración de baterías presentadas en la tabla 1.7 de la página 23 se puede aproximar mediante una distribución normal con media $\mu = 3.5$ y desviación estándar $\sigma = 0.7$. Las frecuencias esperadas para las 7 clases (celdas) que se listan en la tabla 10.5 se obtienen calculando las áreas bajo la curva normal hipotética que caen entre los diversos límites de clase.

Tabla 10.5: Frecuencias observadas y esperadas para la duración de las baterías suponiendo normalidad

Límites de clase	o_i		e_i	
1.45 – 1.95	2	7	0.5	8.5
1.95 – 2.45	1		2.1	
2.45 – 2.95	4		5.9	
2.95 – 3.45	15		10.3	
3.45 – 3.95	10		10.7	
3.95 – 4.45	5	8	7.0	10.5
4.45 – 4.95	3		3.5	

Por ejemplo, los valores z que corresponden a los límites de la cuarta clase son

$$z_1 = \frac{2.95 - 3.5}{0.7} = -0.79 \quad \text{y} \quad z_2 = \frac{3.45 - 3.5}{0.7} = -0.07.$$

En la tabla A.3 encontramos que el área entre $z_1 = -0.79$ y $z_2 = -0.07$ es

$$\begin{aligned} \text{área} = P(-0.79 < Z < -0.07) &= P(Z < -0.07) - P(Z < -0.79) \\ &= 0.4721 - 0.2148 = 0.2573. \end{aligned}$$

Por lo tanto, la frecuencia esperada para la cuarta clase es

$$e_4 = (0.2573)(40) = 10.3.$$

Se acostumbra redondear estas frecuencias a un decimal.

La frecuencia esperada para el primer intervalo de clase se obtiene utilizando el área total bajo la curva normal a la izquierda del límite 1.95. Para el último intervalo de clase usamos el área total a la derecha del límite 4.45. Todas las demás frecuencias esperadas se determinan utilizando el método que se describe para la cuarta clase. Observe que combinamos clases adyacentes en la tabla 10.5 donde las frecuencias esperadas son menores que 5 (una regla general en la prueba de la bondad de ajuste). En consecuencia, el número total de intervalos se reduce de 7 a 4, lo cual da como resultado $v = 3$ grados de libertad. Entonces, el valor χ^2 es dado por

$$\chi^2 = \frac{(7-8.5)^2}{8.5} + \frac{(15-10.3)^2}{10.3} + \frac{(10-10.7)^2}{10.7} + \frac{(8-10.5)^2}{10.5} = 3.05.$$

Como el valor χ^2 calculado es menor que $\chi^2_{0.05} = 7.815$ para 3 grados de libertad, no tenemos razón para rechazar la hipótesis nula y concluimos que la distribución normal con $\mu = 3.5$ y $\sigma = 0.7$ proporciona un buen ajuste para la distribución de la duración de las baterías.

La prueba de bondad de ajuste chi cuadrada es un recurso importante, en particular debido a que muchos procedimientos estadísticos en la práctica dependen, en un sentido teórico, de la suposición de que los datos reunidos provienen de un tipo de distribución específico. Como ya se expuso, la suposición de normalidad se hace muy a menudo. En los siguientes capítulos continuaremos haciendo suposiciones de normalidad con el fin de proporcionar una base teórica para ciertas pruebas e intervalos de confianza.

En la literatura hay pruebas para evaluar la normalidad que son más poderosas que la prueba chi cuadrada. Una de tales pruebas es la **prueba de Geary**, la cual se basa en un estadístico muy sencillo que es el cociente de dos estimadores de la desviación estándar de la población σ. Suponga que se toma una muestra aleatoria $X_1, X_2, \ldots, X_n$ de una distribución normal, $N(\mu, \sigma)$. Considere el cociente

$$U = \frac{\sqrt{\pi/2}\sum_{i=1}^{n} |X_i - \bar{X}|/n}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2/n}}.$$

El lector debería reconocer que el denominador es un estimador razonable de σ sin importar si la distribución es normal o no. El numerador es un buen estimador de σ si la distribución es normal, pero podría sobrestimar o subestimar a σ cuando haya desviaciones de la normalidad. Así, los valores de U que difieren considerablemente de 1.0 representan la señal de que se debe rechazar la hipótesis de normalidad.

Para muestras grandes una prueba razonable se basa en la normalidad aproximada de U. El estadístico de prueba es, entonces, una estandarización de U dada por

$$Z = \frac{U - 1}{0.2661/\sqrt{n}}.$$

Desde luego, el procedimiento de prueba incluye la región crítica bilateral. Calculamos un valor de z a partir de los datos y no rechazamos la hipótesis de normalidad cuando

$$-z_{\alpha/2} < Z < z_{\alpha/2}.$$

En la bibliografía se cita un artículo que trata sobre la prueba de Geary (Geary, 1947).

10.7. Prueba de la independencia (datos categóricos)

El procedimiento de prueba de chi cuadrada que se presentó en la sección 10.11 también se puede usar para probar la hipótesis de independencia de dos variables de clasificación. Suponga que deseamos determinar si las opiniones de los votantes residentes del estado de Illinois respecto a una nueva reforma fiscal son independientes de sus niveles de ingreso. Los sujetos de una muestra aleatoria de 1000 votantes registrados del estado de Illinois se clasifican de acuerdo con su posición en las categorías de ingreso bajo, medio o alto, y si están a favor o no de la nueva reforma fiscal. Las frecuencias observadas se presentan en la tabla 10.6, la cual se conoce como **tabla de contingencia**.

Tabla 10.6: Tabla de contingencia 2×3

	Nivel de ingreso			
Reforma fiscal	**Bajo**	**Medio**	**Alto**	**Total**
A favor	182	213	203	598
En contra	154	138	110	402
Total	336	351	313	1000

Una tabla de contingencia con r renglones y c columnas se denomina tabla $r \times c$ ("$r \times c$" se lee "r por c"). Los totales de renglones y columnas en la tabla 10.6 se denominan **frecuencias marginales**. Nuestra decisión de aceptar o rechazar la hipótesis nula, H_0, de que la opinión de un votante respecto a la nueva reforma fiscal es independiente de su nivel de ingreso, se basa en qué tan bien se ajusten las frecuencias observadas en cada una de las 6 celdas de la tabla 10.6 y en las frecuencias que esperaríamos para cada celda si supusiéramos que H_0 es verdadera. Para encontrar estas frecuencias esperadas definamos los siguientes eventos:

L: Una persona seleccionada está en el nivel de ingresos bajo.

M: Una persona seleccionada está en el nivel de ingresos medio.

H: Una persona seleccionada está en el nivel de ingresos alto.

F: Una persona seleccionada está a favor de la nueva reforma fiscal.

A: Una persona seleccionada está en contra de la nueva reforma fiscal.

Podemos usar las frecuencias marginales para listar las siguientes estimaciones de probabilidad:

$$P(L) = \frac{336}{1000}, \quad P(M) = \frac{351}{1000}, \quad P(H) = \frac{313}{1000},$$
$$P(F) = \frac{598}{1000}, \quad P(A) = \frac{402}{1000}.$$

Ahora bien, si H_0 es verdadera y las dos variables son independientes, deberíamos tener

$$P(L \cap F) = P(L)\,P(F) = \left(\frac{336}{1000}\right)\left(\frac{598}{1000}\right),$$
$$P(L \cap A) = P(L)\,P(A) = \left(\frac{336}{1000}\right)\left(\frac{402}{1000}\right),$$

$$P(M \cap F) = P(M)\,P(F) = \left(\frac{351}{1000}\right)\left(\frac{598}{1000}\right),$$
$$P(M \cap A) = P(M)\,P(A) = \left(\frac{351}{1000}\right)\left(\frac{402}{1000}\right),$$
$$P(H \cap F) = P(H)\,P(F) = \left(\frac{313}{1000}\right)\left(\frac{598}{1000}\right),$$
$$P(H \cap A) = P(H)\,P(A) = \left(\frac{313}{1000}\right)\left(\frac{402}{1000}\right).$$

Las frecuencias esperadas se obtienen multiplicando la probabilidad de cada celda por el número total de observaciones. Como antes, redondeamos estas frecuencias a un decimal. Así, se estima que el número esperado de votantes de bajo ingreso en nuestra muestra que favorecen la reforma fiscal es

$$\left(\frac{336}{1000}\right)\left(\frac{598}{1000}\right)(1000) = \frac{(336)(598)}{1000} = 200.9$$

cuando H_0 es verdadera. La regla general para obtener la frecuencia esperada de cualquier celda es dada por la siguiente fórmula:

$$\text{frecuencia esperada} = \frac{(\text{total por columna}) \times (\text{total por renglón})}{\text{gran total}}$$

En la tabla 10.7 la frecuencia esperada para cada celda se registra entre paréntesis, a un lado del valor observado verdadero. Observe que las frecuencias esperadas en cualquier renglón o columna se suman al total marginal apropiado. En nuestro ejemplo necesitamos calcular sólo las dos frecuencias esperadas en el renglón superior de la tabla 10.7 y luego calcular las otras mediante sustracción. El número de grados de libertad asociados con la prueba chi cuadrada que aquí se usa es igual al número de frecuencias de celdas que se pueden llenar libremente cuando se nos proporcionan los totales marginales y el gran total, y en este caso ese número es 2. Una fórmula sencilla que proporciona el número correcto de grados de libertad es

$$v = (r - 1)(c - 1).$$

Tabla 10.7: Frecuencias observadas y esperadas

	Nivel de ingreso			
Reforma fiscal	**Bajo**	**Medio**	**Alto**	**Total**
A favor	182 (200.9)	213 (209.9)	203 (187.2)	598
En contra	154 (135.1)	138 (141.1)	110 (125.8)	402
Total	336	351	313	1000

Por lo tanto, para nuestro ejemplo $v = (2 - 1)(3 - 1) = 2$ grados de libertad. Para probar la hipótesis nula de independencia usamos el siguiente criterio de decisión:

Prueba de independencia

Calcule

$$\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i},$$

donde la sumatoria se extiende a todas las celdas rc en la tabla de contingencia $r \times c$.

Si $\chi^2 > \chi^2_\alpha$ con $v = (r-1)(c-1)$ grados de libertad, rechace la hipótesis nula de independencia al nivel de significancia α; en otro caso no la rechace.

Al aplicar este criterio a nuestro ejemplo encontramos que

$$\chi^2 = \frac{(182-200.9)^2}{200.9} + \frac{(213-209.9)^2}{209.9} + \frac{(203-187.2)^2}{187.2} + \frac{(154-135.1)^2}{135.1} + \frac{(138-141.1)^2}{141.1} + \frac{(110-125.8)^2}{125.8} = 7.85,$$

$$P \approx 0.02.$$

En la tabla A.5 encontramos que $\chi^2_{0.05} = 5.991$ para $v = (2-1)(3-1) = 2$ grados de libertad. Rechazamos la hipótesis nula y concluimos que la opinión de un votante respecto a la reforma fiscal y su nivel de ingresos no son independientes.

Es importante recordar que el estadístico sobre el cual basamos nuestra decisión tiene una distribución que sólo se aproxima por la distribución chi cuadrada. Los valores χ^2 calculados dependen de las frecuencias de las celdas y, en consecuencia, son discretos. La distribución chi cuadrada continua parece aproximarse muy bien a la distribución de muestreo discreta de χ^2, siempre y cuando el número de grados de libertad sea mayor que 1. En una tabla de contingencia de 2×2, donde sólo tenemos 1 grado de libertad, se aplica una corrección llamada **corrección de Yates para continuidad**.

La fórmula corregida entonces se convierte en

$$\chi^2(\text{corregida}) = \sum_i \frac{(|o_i - e_i| - 0.5)^2}{e_i}.$$

Si las frecuencias de las celdas esperadas son grandes, los resultados corregidos y sin corrección son casi iguales. Cuando las frecuencias esperadas están entre 5 y 10, se debe aplicar la corrección de Yates. Para frecuencias esperadas menores que 5 se debería utilizar la prueba exacta de Fisher-Irwin. Un análisis de esta prueba se puede encontrar en *Basic Concepts of Probability and Statistics* de Hodges y Lehmann (2005; véase la bibliografía). Sin embargo, la prueba de Fisher-Irwin se puede evitar seleccionando una muestra grande.

10.8. Prueba de Homogeneidad

Cuando probamos la independencia en la sección 10.12 seleccionamos una muestra aleatoria de 1000 votantes, y determinamos al azar los totales de renglón y de columna para nuestra tabla de contingencia. Otro tipo de problema para el que se aplica el método de la sección 10.12 es aquel en el cual los totales de renglón y de columna están predeterminados. Suponga, por ejemplo, que decidimos de antemano seleccionar 200 demócratas, 150 republicanos y 150 independientes entre los votantes del estado de Carolina del Norte y registrar si están a favor de una iniciativa de ley para el aborto, si están en contra o si están indecisos. Las respuestas observadas se incluyen en la tabla 10.8.

Tabla 10.8: Frecuencias observadas

	Afiliación política			
Ley para el aborto	**Demócrata**	**Republicano**	**Independiente**	**Total**
A favor	82	70	62	214
En contra	93	62	67	222
Indeciso	25	18	21	64
Total	200	150	150	500

Ahora bien, en vez de hacer una prueba de independencia, probamos la hipótesis de que las proporciones de población dentro de cada renglón son iguales. Es decir, probamos la hipótesis de que las proporciones de demócratas, republicanos e independientes que están a favor de la ley para el aborto son iguales; las proporciones de cada afiliación política contra la ley son iguales y las proporciones de cada afiliación política que están indecisos son iguales. Básicamente nos interesamos en determinar si las tres categorías de votantes son **homogéneas** en lo que se refiere a sus opiniones acerca de la iniciativa de ley para el aborto. A esta prueba se le conoce como prueba de homogeneidad.

Al suponer homogeneidad de nuevo calculamos las frecuencias esperadas de las celdas multiplicando los totales de renglón y de columna correspondientes y después dividiendo entre el gran total. Luego continuamos el análisis utilizando el mismo estadístico chi cuadrada como antes. Ilustramos este proceso en el siguiente ejemplo para los datos de la tabla 10.8.

11. REGRESIÓN LINEAL SIMPLE Y CORRELACIÓN

11.1.Introducción a la regresión lineal

En la práctica a menudo se requiere resolver problemas que implican conjuntos de variables de las cuales se sabe que tienen alguna relación inherente entre sí. Por ejemplo, en una situación industrial quizá se sepa que el contenido de alquitrán en el flujo de salida de un proceso químico está relacionado con la temperatura en la entrada. Podría ser de interés desarrollar un método de pronóstico, es decir, un procedimiento que permita estimar el contenido de alquitrán para varios niveles de temperatura de entrada a partir de información experimental. Desde luego, es muy probable que para muchos ejemplos concretos en los que la temperatura de entrada sea la misma, por ejemplo 130ºC, el contenido de alquitrán de salida no sea el mismo. Esto es muy similar a lo que ocurre cuando se estudian varios automóviles con un motor del mismo volumen; no todos tienen el mismo rendimiento de combustible. No todas las casas ubicadas en la misma zona del país, con la misma superficie de construcción, se venden al mismo precio. El contenido de alquitrán, el rendimiento del combustible (en millas por galón) y el precio de las casas (en miles de dólares) son **variables dependientes** naturales o respuestas en los tres escenarios. La temperatura en la entrada, el volumen del motor (pies cúbicos) y los metros cuadrados de superficie de construcción son, respectivamente, **variables independientes** naturales o **regresores**. Una forma razonable de relación entre la **respuesta** Y y el regresor x es la relación lineal,

$$Y = \beta_0 + \beta_1 x,$$

en la que, por supuesto, β_0 es la **intersección** y β_1 es la **pendiente**. Esta relación se ilustra en la figura 11.1.

Si la relación es exacta y no contiene ningún componente aleatorio o probabilístico, entonces se trata de una relación **determinista** entre dos variables científicas. Sin embargo, en los ejemplos que se mencionaron, así como en muchos otros fenómenos científicos y de ingeniería, la relación no es determinista, es decir, una x dada no siempre produce el mismo valor de Y. Como resultado, los problemas importantes en este caso son de naturaleza probabilística, toda vez que la relación anterior no puede considerarse exacta. El concepto de **análisis de regresión** se refiere a encontrar la mejor relación entre Y y x

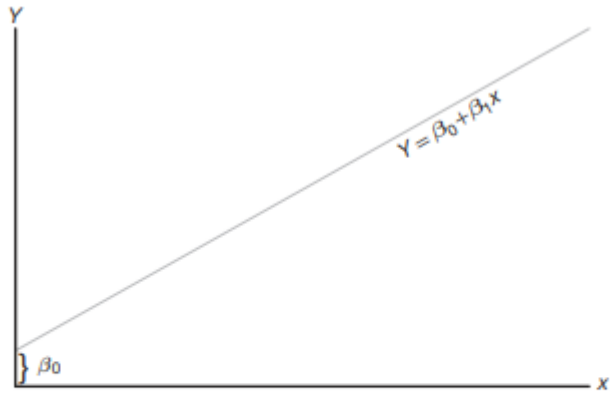


Figura 11.1: Una relación lineal; β_0: intersección; β_1: pendiente.

cuantificando la fuerza de esa relación, y empleando métodos que permitan predecir los valores de la respuesta dados los valores del regresor x.

En muchas aplicaciones habrá más de un regresor, es decir, más de una variable independiente **que ayude a explicar a *Y***. Por ejemplo, si se tratara de explicar las razones para el precio de una casa, se esperaría que una de ellas fuera su antigüedad, en cuyo caso la estructura múltiple de la regresión se podría escribir como

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

donde Y es el precio, x_1 son los metros cuadrados y x_2 es la antigüedad de la casa en años. En el capítulo siguiente se estudiarán problemas con regresores múltiples. El análisis resultante se denomina **regresión múltiple**; en tanto que el análisis del caso con un solo regresor recibe el nombre de **regresión simple**. En un segundo ejemplo de la regresión múltiple, un ingeniero químico podría estar interesado en la cantidad de hidrógeno que se ha perdido en las muestras de un metal específico que se tiene almacenado. En este caso habría dos entradas, x_1, el tiempo de almacenamiento en horas, y x_2, la temperatura de almacenamiento en grados centígrados. De modo que la respuesta sería Y, la pérdida de hidrógeno en partes por millón.

En este capítulo estudiaremos el tema de la **regresión lineal simple**, que trata el caso de una sola variable regresora, en el que la relación entre x y y es lineal. Para el caso en el que hay más de una variable regresora el lector debe consultar el capítulo 12. Denotemos una muestra aleatoria de tamaño n mediante el conjunto $\{(x_i, y_i); i = 1, 2,..., n\}$. Si se tomaran muestras adicionales utilizando exactamente los mismos valores de x, se esperaría que los valores de y variaran. Así, el valor y_i en el par ordenado (x_i, y_i) es el valor de cierta variable aleatoria Y_i.

11.2. regresión lineal simple

Hemos limitado el uso del término *análisis de regresión* a los casos en los que las relaciones entre las variables no son deterministas, es decir, no son exactas. En otras palabras, debe existir un **componente aleatorio** en la ecuación que relaciona las variables. Este componente aleatorio toma en cuenta consideraciones que no son medibles o, de hecho, que los científicos o los ingenieros no comprenden. En realidad, en la mayoría de aplicaciones de la regresión, la ecuación lineal, digamos, $Y = \beta_0 + \beta_1 x$ es una aproximación que representa de manera simplificada algo desconocido y mucho más complicado. Por ejemplo, en el caso que implica la respuesta Y = contenido de alquitrán y x = temperatura de entrada es probable que $Y = \beta_0 + \beta_1 x$ sea una aproximación razonable que podría funcionar dentro de un rango limitado de x. La mayoría de las veces los modelos que son simplificaciones de estructuras más complicadas y desconocidas son de naturaleza lineal, es decir, lineales en los **parámetros** β_0 y β_1 o, en el caso del modelo que implica el precio, el tamaño y la antigüedad de la casa, lineal en los **parámetros** β_0, β_1 y β_2. Estas estructuras lineales son sencillas y de naturaleza empírica, por lo que se denominan **modelos empíricos**.

Un análisis de la relación entre x y Y requiere el planteamiento de un **modelo estadístico**. Con frecuencia un estadístico utiliza un modelo como representación de un **ideal** que, en esencia, define cómo percibimos que el sistema en cuestión generó los datos. El modelo debe incluir al conjunto $\{(x_i, y_i); i = 1, 2,..., n\}$ de datos que implica n pares de valores (x, y). No debemos olvidar que el valor de y_i depende de x_i por medio de una estructura lineal que también incluye el componente aleatorio. La base para el uso de un modelo estadístico se relaciona con la manera en que la variable aleatoria Y cambia con x y el componente aleatorio. El modelo también incluye lo que se asume acerca de las propiedades estadísticas del componente aleatorio. A continuación se presenta el modelo estadístico para la regresión lineal simple. La respuesta Y se relaciona con la variable independiente x a través de la ecuación

Modelo de regresión lineal simple

$$Y = \beta_0 + \beta_1 x + \epsilon.$$

en la cual β_0 y β_1 son los parámetros desconocidos de la intersección y la pendiente, respectivamente, y ϵ es una variable aleatoria que se supone está distribuida con $E(\epsilon) = 0$ y $\text{Var}(\epsilon) = \sigma^2$. Es frecuente que a la cantidad σ^2 se le denomine varianza del error o varianza residual.

En el modelo anterior hay varias cuestiones evidentes. La cantidad Y es una variable aleatoria, ya que ϵ es aleatoria. El valor x de la variable regresora no es aleatorio y, de hecho, se mide con un error despreciable. La cantidad ϵ, que a menudo recibe el nombre de **error aleatorio** o **alteración aleatoria**, tiene varianza constante. Es común que a esta parte se le denomine **suposición de varianza homogénea**. La presencia de este error aleatorio ϵ evita que el modelo se convierta tan sólo en una ecuación determinista. Ahora, el hecho de que $E(\epsilon) = 0$ implica que para una x específica, los valores de y se distribuyen alrededor de la **recta verdadera** o **recta de regresión** de la población $y = \beta_0 + \beta_1 x$. Si se elige bien el modelo, es decir, si no hay otros regresores de importancia y la aproximación lineal es buena dentro de los rangos de los datos, entonces son razonables los errores positivos y negativos que rodean a la regresión verdadera. Debe recordarse que en la práctica β_0 y β_1 se desconocen y que deben estimarse a partir de los datos. Además, el modelo que se acaba de describir es de naturaleza conceptual. Como resultado, en la práctica nunca se observan los valores ϵ reales, por lo que nunca se puede trazar la verdadera recta de regresión, aunque suponemos que ahí está. Sólo es posible dibujar una recta estimada. En la figura 11.2 se ilustra la naturaleza de los datos (x, y) hipotéticos dispersos alrededor de la verdadera recta de regresión para un caso en que

sólo se dispone de $n = 5$ observaciones. Debemos destacar que lo que observamos en la figura 11.2 no es la recta que utilizan el científico o ingeniero. En vez de esa recta, ¡lo

que describe la ilustración es el significado de las suposiciones! Ahora describiremos la regresión que el usuario tiene a su disposición.

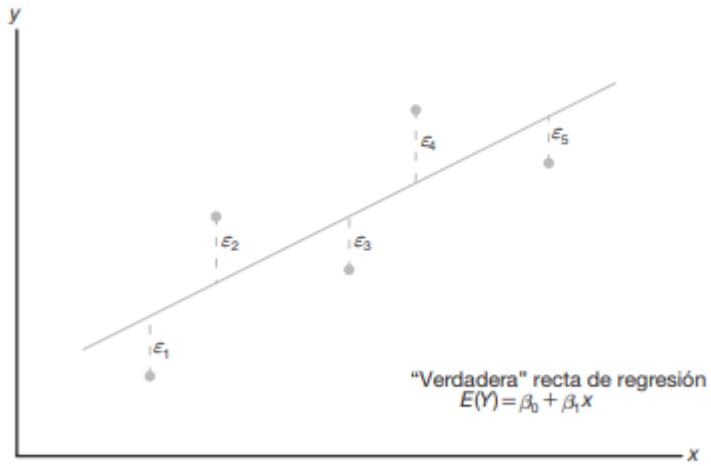


Figura 11.2: Datos (x, y) hipotéticos dispersos alrededor de la verdadera recta de regresión para $n = 5$.

La recta de regresión ajustada

Un aspecto importante del análisis de regresión es, en términos sencillos, estimar los parámetros β_0 y β_1, es decir, estimar los llamados **coeficientes de regresión**. En la sección siguiente se estudiará el método para estimarlos. Suponga que denotamos los estimados b_0 para β_0 y b_1 para β_1. Entonces, la recta de **regresión ajustada**, o estimada, es dada por

$$\hat{y} = b_0 + b_1 x,$$

donde $\hat{y}$ es el valor pronosticado o ajustado. Es evidente que la recta ajustada es un estimado de la verdadera recta de regresión. Se espera que la recta ajustada esté más cerca de la verdadera línea de regresión cuando se dispone de una gran cantidad de datos. En el ejemplo siguiente se ilustra la recta ajustada para un estudio sobre contaminación en la vida real.

Uno de los problemas más desafiantes que enfrenta el campo del control de la contaminación del agua lo representa la industria de la peletería, ya que sus desechos son químicamente complejos; se caracterizan por valores elevados de la demanda de oxígeno químico, sólidos volátiles y otras medidas de contaminación. Considere los datos experimentales de la tabla 11.1, que se obtuvieron de 33 muestras de desechos tratados químicamente en un estudio realizado en Virginia Tech. Se registraron los valores de x, la reducción porcentual de los sólidos totales, y de y, el porcentaje de disminución de la demanda de oxígeno químico.

Otra mirada a las suposiciones del modelo

Resulta aleccionador repasar el modelo de regresión lineal simple que se presentó con anterioridad y analizar de forma gráfica la manera en que se relaciona con la denominada regresión verdadera. Daremos más detalles en la figura 11.2, cuando ilustremos no sólo el lugar en que los ϵ_i se localizan en la gráfica, sino también lo que implica la suposición de normalidad para los ϵ_i.

Suponga que tenemos una regresión lineal simple con $n = 6$, valores de x equidistantes y un valor único de y para cada x. Considere la gráfica de la figura 11.4, la cual debería proporcionar al lector una representación clara del modelo y de las suposiciones implicadas. La recta que aparece en la gráfica es la recta de regresión verdadera. Los puntos graficados (y, x) son puntos reales dispersos alrededor de la recta. Cada punto se ubica en su propia distribución normal, donde el centro de la distribución, es decir, la media de y, cae sobre la recta. Ciertamente esto es lo esperado, ya que $E(Y) = \beta_0 + \beta_1 x$. Como resultado, la verdadera recta de regresión **pasa a través de las medias de la respuesta** y las observaciones reales se encuentran sobre la distribución, alrededor de las medias. Observe también que todas las distribuciones tienen la misma varianza, que se denota con σ^2. Desde luego, la desviación entre una y individual y el punto sobre la recta será su valor individual ϵ. Esto queda claro porque

$$y_i - E(Y_i) = y_i - (\beta_0 + \beta_1 x_i) = \epsilon_i.$$

Así, con una x dada, tanto Y como el ϵ correspondiente tienen varianza σ^2.

Así, con una x dada, tanto Y como el ϵ correspondiente tienen varianza σ^2.

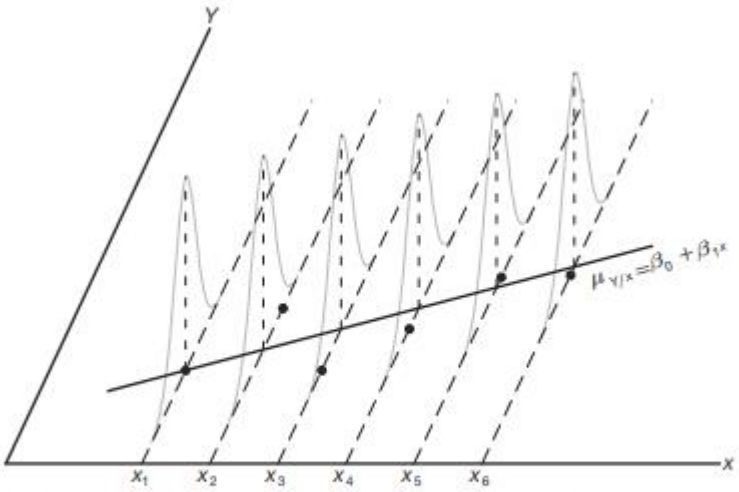


Figura 11.4: Observaciones individuales alrededor de la verdadera recta de regresión.

Note también que aquí escribimos la verdadera recta de regresión como $\mu_{Y|x} = \beta_0 + \beta_1 x$ con el fin de reafirmar que la recta pasa a través de la media de la variable aleatoria Y.

11.3. Propiedades de los estimadores de mínimos cuadrados

Mínimos cuadrados y el modelo ajustado

En esta sección se estudia el método para ajustar una recta de regresión estimada a los datos, lo cual equivale a determinar los estimados b_0 para β_0 y b_1 para β_1. Por supuesto, esto permite el cálculo de los valores pronosticados a partir de la recta ajustada $\hat{y} = b_0 + b_1x$, y otros tipos de análisis y de información diagnóstica que determinarán la fuerza de la relación, así como la adecuación y el ajuste del modelo. Antes de analizar el método de estimación de los mínimos cuadrados es importante presentar el concepto de **residual**. En esencia, un residual es un error en el ajuste del modelo $\hat{y} = b_0 + b_1x$.

Residual: Error en el ajuste

Dado un conjunto de datos de regresión $\{(x_i, y_i); i = 1, 2,..., n\}$ y un modelo ajustado $\hat{y}_i = b_0 + b_1x$, el i-ésimo residual e_i es dado por

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \ldots, n.$$

Es evidente que si un conjunto de n residuales es grande, entonces el ajuste del modelo no es bueno. Los residuales pequeños son indicadores de un ajuste adecuado. Otra relación interesante, y que a veces es útil, es la siguiente:

$$y_i = b_0 + b_1x_i + e_i.$$

El uso de la ecuación anterior debería aclarar la diferencia entre los residuales e_i y los errores del modelo conceptual ϵ_i. No debemos olvidar que, mientras que los ϵ_i no se observan, los e_i no sólo se observan sino que desempeñan un papel importante en el análisis total.

La figura 11.5 ilustra el ajuste de la recta a este conjunto de datos: a saber $\hat{y} = b_0 + b_1x$, y la recta que refleja el modelo $\mu_{Y|x} = \beta_0 + \beta_1x$. Desde luego, β_0 y β_1 son parámetros desconocidos. La recta ajustada es un estimado de la recta que genera el modelo estadístico. Hay que tener presente que la recta $\mu_{Y|x} = \beta_0 + \beta_1x$ es desconocida.

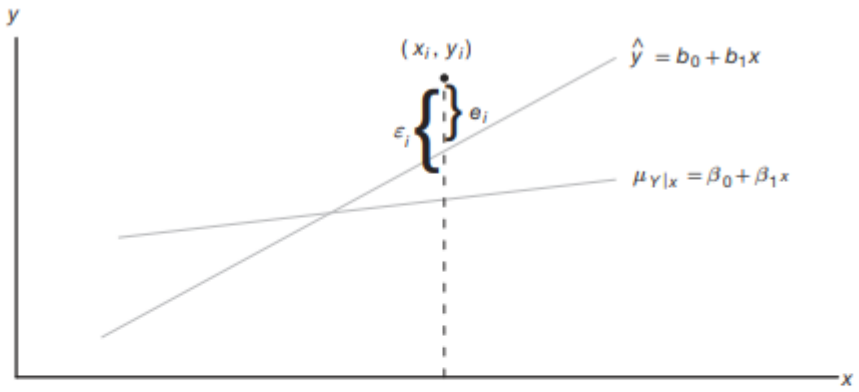


Figura 11.5: Comparación de ϵ_i con el residual e_i.

Método de mínimos cuadrados

Debemos calcular b_0 y b_1, los estimados de β_0 y β_1, de manera que la suma de los cuadrados de los residuales sea mínima. La suma residual de los cuadrados con frecuencia se denomina suma de los cuadrados del error respecto de la recta de regresión y se denota como *SCE*. Este procedimiento de minimización para estimar los parámetros se denomina **método de mínimos cuadrados**. Por lo tanto, debemos calcular a y b para minimizar

$$SCE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2.$$

Al diferenciar la *SCE* con respecto a b_0 y b_1, se obtiene

$$\frac{\partial(SCE)}{\partial b_0} = -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i), \quad \frac{\partial(SCE)}{\partial b_1} = -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)x_i.$$

Al igualar a cero las derivadas parciales y reacomodar los términos, obtenemos las ecuaciones siguientes (llamadas **ecuaciones normales**)

$$nb_0 + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \quad b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i,$$

que se resuelven simultáneamente para obtener fórmulas de cálculo para b_0 y b_1.

Estimación de los coeficientes de regresión

Dada la muestra $\{(x_i, y_i)\}; i = 1, 2, \ldots, n\}$, los estimados b_0 y b_1 de los mínimos cuadrados de los coeficientes de regresión β_0 y β_1 se calculan mediante las fórmulas

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \text{ y}$$

$$b_0 = \frac{\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i}{n} = \bar{y} - b_1\bar{x}.$$

Media y varianza de los estimadores

En la exposición que sigue mostramos que el estimador B_1 es insesgado para β_1, y se demuestran tanto las varianzas de B_0 como las de B_1. Esto inicia una serie de procedimientos que conducen a la prueba de hipótesis y a la estimación de intervalos de confianza para la intersección y la pendiente.

Como el estimador

$$B_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

es de la forma $\sum_{i=1}^{n} c_i Y_i$,

$$c_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad i = 1, 2, \ldots, n,$$

podemos concluir a partir del teorema 7.11 que B_1 tiene una distribución $n(\mu_{B_1}, \sigma_{B_1})$ con

$$\mu_{B_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \beta_1 \text{ y } \sigma^2_{B_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sigma^2_{Y_i}}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

También se puede demostrar (véase el ejercicio de repaso 11.60 de la página 438) que la variable aleatoria B_0 se distribuye normalmente con

$$\text{media } \mu_{B_0} = \beta_0 \text{ y varianza } \sigma^2_{B_0} = \frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}\sigma^2.$$

A partir de estos resultados es evidente que los **estimadores de mínimos cuadrados tanto para β_0 como para β_1 son insesgados**.

Partición de la variabilidad total y estimación de σ^2

Para hacer inferencias sobre β_0 y β_1 es necesario llegar a una estimación del parámetro σ^2 que aparece en las dos fórmulas anteriores de la varianza de B_0 y B_1. El parámetro σ^2, el modelo de la varianza del error, refleja una variación aleatoria o una variación del

error experimental alrededor de la recta de regresión. En gran parte de lo que sigue se recomienda emplear la notación

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2, \quad S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2, \quad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

De manera que la suma de los cuadrados del error se puede escribir como sigue:

$$\begin{aligned} SCE &= \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2 = \sum_{i=1}^{n}[(y_i - \bar{y}) - b_1(x_i - \bar{x})]^2 \\ &= \sum_{i=1}^{n}(y_i - \bar{y})^2 - 2b_1 \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) + b_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= S_{yy} - 2b_1 S_{xy} + b_1^2 S_{xx} = S_{yy} - b_1 S_{xy}, \end{aligned}$$

que es el paso final que surge del hecho de que $b_1 = S_{xy}/S_{xx}$.

Teorema 11.1: Un estimador insesgado de σ^2 es

$$s^2 = \frac{SCE}{n-2} = \sum_{i=1}^{n}\frac{(y_i - \hat{y}_i)^2}{n-2} = \frac{S_{yy} - b_1 S_{xy}}{n-2}.$$

El estimador de σ^2 como error cuadrado medio

Para darnos una idea del estimador de σ^2 deberíamos observar el resultado del teorema 11.1. El parámetro σ^2 mide la varianza o las desviaciones cuadradas entre los valores de Y y su media, dada por $\mu_{Y|x}$, es decir, las desviaciones cuadradas entre Y y $\beta_0 + \beta_1 x$. Por supuesto, $\beta_0 + \beta_1 x$ se estima por medio de $\hat{y} = b_0 + b_1 x$. Por consiguiente, tendría sentido que la varianza σ^2 se describa mejor como una desviación cuadrada de la observación típica y_i con respecto a la media estimada $\hat{y}_i$, que es el punto correspondiente sobre la recta ajustada. Entonces, los valores $(y_i - \hat{y}_i)$ revelan la varianza apropiada, de manera muy similar a como los valores $(y_i - \bar{y})^2$ miden la varianza cuando se realiza un muestreo en un escenario no relacionado con la regresión. En otras palabras, $\bar{y}$ estima la media en la última situación sencilla, mientras que $\hat{y}_i$ estima la media de y_i en una estructura de regresión. Ahora, ¿qué significa el divisor $n - 2$? En las secciones que siguen observaremos que éstos son los grados de libertad asociados con el estimador s^2 de σ^2. En tanto que en el escenario i.i.d. (independiente e idénticamente distribuidas), la normal estándar se resta un grado de libertad de n en el denominador, para lo cual una explicación razonable es que se estima un parámetro, que es la media μ por medio de, digamos, $\bar{y}$, pero en el problema de la regresión **se estiman dos parámetros**, que son β_0 y β_1, por medio de b_0 y b_1. Así, el parámetro importante σ^2, que se estima mediante

$$s^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-2),$$

se denomina **error cuadrado medio**, que describe un tipo de media (división entre $n - 2$) de los residuales cuadrados.

11.4. inferencias acerca de los coeficientes de regresión

Además de tan sólo estimar la relación lineal entre x y Y para fines de predicción, el experimentador podría estar interesado en hacer ciertas inferencias acerca de la pendiente y la intersección. Para dar ocasión a la prueba de hipótesis y a la construcción de intervalos de confianza para β_0 y β_1, debemos estar dispuestos a hacer la suposición adicional de que cada ϵ_i, $i = 1, 2, ..., n$, se distribuye de forma normal. Esta suposición implica que $Y_1, Y_2, ..., Y_n$ también están distribuidas normalmente, cada una con una distribución de probabilidad $n(y_i; \beta_0 + \beta_1 x_i, \sigma)$.

A partir de la sección 11.4 sabemos que B_1 tiene una distribución normal, y suponiendo normalidad, un resultado muy parecido al que se plantea en el teorema 8.4 nos permite concluir que $(n-2)S^2/\sigma^2$ es una variable chi cuadrada con $n-2$ grados de libertad, independiente de la variable aleatoria B_1. Entonces, el teorema 8.5 garantiza que el estadístico

$$T = \frac{(B_1 - \beta_1)/(\sigma/\sqrt{S_{xx}})}{S/\sigma} = \frac{B_1 - \beta_1}{S/\sqrt{S_{xx}}}$$

tenga una distribución t con $n-2$ grados de libertad. Podemos utilizar el estadístico T para construir un intervalo de confianza del $100(1-\alpha)\%$ para el coeficiente β_1.

Intervalo de confianza para β_1

Un intervalo de confianza de $100(1-\alpha)\%$ para el parámetro β_1 en la recta de regresión $\mu_{Y|x} = \beta_0 + \beta_1 x$ es

$$b_1 - t_{\alpha/2}\frac{s}{\sqrt{S_{xx}}} < \beta_1 < b_1 + t_{\alpha/2}\frac{s}{\sqrt{S_{xx}}},$$

donde $t_{\alpha/2}$ es un valor de la distribución t con $n-2$ grados de libertad.

Prueba de hipótesis sobre la pendiente

Para probar la hipótesis nula H_0 de que $\beta_1 = \beta_{10}$, en comparación con una alternativa posible, utilizamos de nuevo la distribución t con $n-2$ grados de libertad con el fin de establecer una región crítica y después basar nuestra decisión en el valor de

$$t = \frac{b_1 - \beta_{10}}{s/\sqrt{S_{xx}}}.$$

El método se ilustra con el ejemplo siguiente.

Inferencia estadística sobre la intersección

Los intervalos de confianza y la prueba de hipótesis del coeficiente β_0 se podrían establecer a partir del hecho de que B_0 también se distribuye de forma normal. No es difícil demostrar que

$$T = \frac{B_0 - \beta_0}{S\sqrt{\sum_{i=1}^{n} x_i^2/(nS_{xx})}}$$

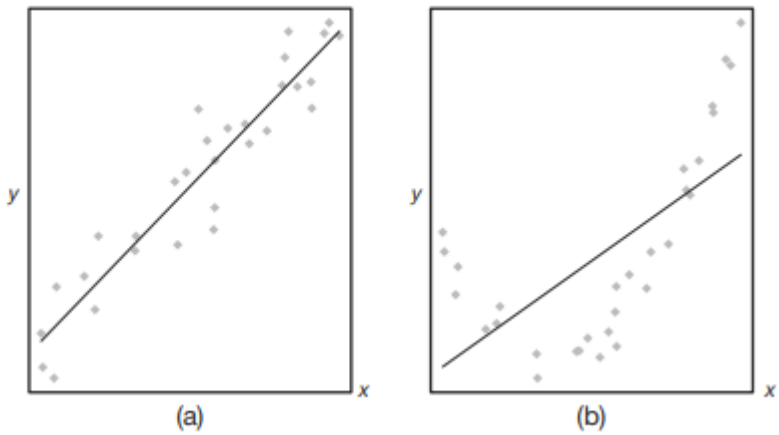


Figura 11.9: Se rechaza la hipótesis de que H_0: $\beta_1 = 0$.

tiene una distribución t con $n - 2$ grados de libertad, de manera que podemos construir un intervalo de confianza de $100(1 - \alpha)\%$ para α.

Intervalo de confianza para β_0

Un intervalo de confianza de $100(1 - \alpha)\%$ para el parámetro β_0 en la recta de regresión $\mu_{Y|x} = \beta_0 + \beta_1 x$ es

$$b_0 - t_{\alpha/2}\frac{s}{\sqrt{nS_{xx}}}\sqrt{\sum_{i=1}^{n} x_i^2} < \beta_0 < b_0 + t_{\alpha/2}\frac{s}{\sqrt{nS_{xx}}}\sqrt{\sum_{i=1}^{n} x_i^2},$$

donde $t_{\alpha/2}$ es un valor de la distribución t con $n - 2$ grados de libertad.

Una medida de la calidad del ajuste: el coeficiente de determinación

Observe en la figura 11.7 que aparece un elemento denotado con R-Sq, cuyo valor es 91.3%. Esta cantidad, R^2, se denomina **coeficiente de determinación** y es una medida de la **proporción de la variabilidad explicada por el modelo ajustado**. En la sección 11.8 se presentará el concepto del método del análisis de varianza para la prueba de hipótesis en la regresión. El enfoque del análisis de varianza utiliza la suma de los cuadrados del error $SCE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ y la **suma total de los cuadrados corregida** $STCC = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$. Esta última representa la variación en los valores de respuesta que *idealmente* serían explicados con el modelo. El valor de la SCE es la variación debida al error, o la **variación no explicada**. Resulta claro que si la $SCE = 0$, toda variación queda explicada. La cantidad que representa la variación explicada es $STCC - SCE$. R^2 es el

$$\text{Coeficiente de determinación: } R^2 = 1 - \frac{SCE}{STCC}.$$

Advierta que si el ajuste es perfecto, *todos los residuales son cero*, y así $R^2 = 1.0$. Pero si la SCE es tan sólo un poco menor que la $STCC$, $R^2 \approx 0$. Observe en la salida de resultados de la figura 11.7 que el coeficiente de determinación sugiere que el modelo ajustado a los datos explica el 91.3% de la variabilidad observada en la respuesta, la reducción en la demanda de oxígeno químico.

La figura 11.10 ofrece ejemplos de una gráfica con un buen ajuste ($R^2 \approx 1.0$) en a) y una gráfica con un ajuste deficiente ($R^2 \approx 0$) en b).

Errores en el uso de R^2

Los analistas citan con mucha frecuencia los valores de R^2, quizá debido a su simplicidad. Sin embargo, hay errores en su interpretación. La confiabilidad de R^2 depende del

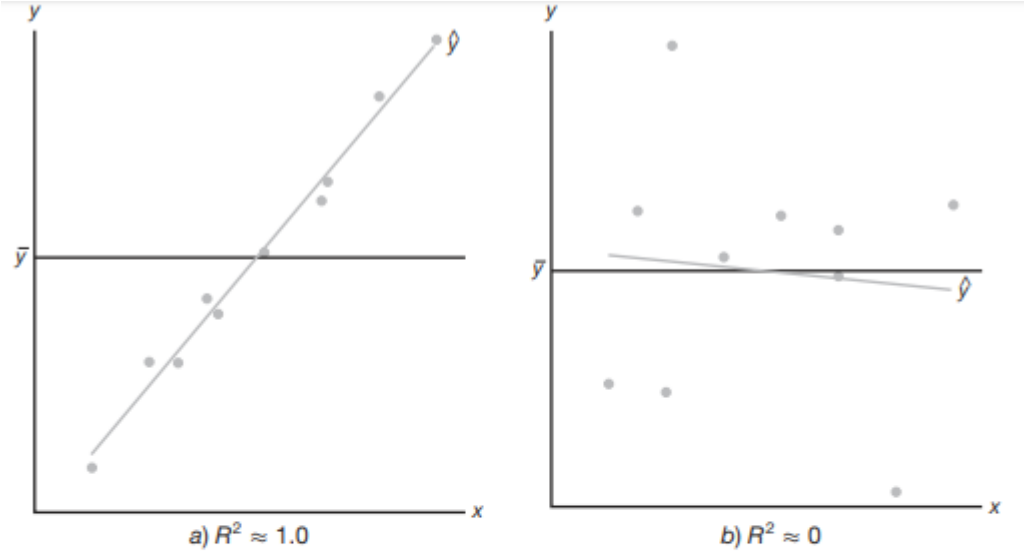


Figura 11.10: Gráficas que ilustran un ajuste muy bueno y otro deficiente.

tamaño del conjunto de los datos de la regresión y del tipo de aplicación. Resulta claro que $0 \leq R^2 \leq 1$, y el límite superior se logra cuando el ajuste a los datos es perfecto, es decir, cuando todos los residuales son cero. ¿Cuál es un valor aceptable de R^2? Se trata de una pregunta difícil de responder. Un químico encargado de establecer una calibración lineal de una pieza de equipo de alta precisión seguramente esperaría obtener un valor muy alto de R^2 (quizá superior a 0.99); mientras que un científico del comportamiento, que trabaja con datos en los que influye la variabilidad de la conducta humana, quizá se sentiría afortunado si obtuviera un valor de R^2 de hasta 0.70. Un individuo con experiencia en el ajuste de modelos tiene la sensibilidad para saber cuándo un valor es suficientemente grande dada la situación que está enfrentando. Es evidente que algunos fenómenos científicos se prestan más a un modelamiento más preciso que otros.

Es peligroso usar el criterio de R^2 para comparar *modelos en competencia* para el mismo conjunto de datos. Cuando se agregan términos adicionales al modelo, por ejemplo un regresor más, disminuye la *SCE*, lo que provoca que R^2 aumente (o al menos no disminuya). Esto implica que R^2 se puede volver artificialmente elevado por medio de la práctica inapropiada de **sobreajustar**, es decir, de incluir demasiados términos en el modelo. Por consiguiente, el incremento inevitable de R^2 que se logra al agregar términos adicionales no implica que éstos se necesitaban. En realidad, el modelo simple puede ser mejor para predecir los valores de la respuesta. En el capítulo 12, cuando se presente el concepto de los modelos que implican **más de un solo regresor**, se estudiará con detalle el papel del sobreajuste y su influencia sobre la capacidad de predicción. En este momento baste decir que *para seleccionar un modelo no se debe adoptar un proceso de selección que sólo incluya la consideración de R^2.*

11.5. Predicción

Hay varias razones para construir un modelo de regresión lineal. Una de ellas es, desde luego, predecir valores de respuesta para uno o más valores de la variable independiente. En esta sección se centra el enfoque en los errores asociados con la predicción.

La ecuación $\hat{y} = b_0$ y b_1x se puede utilizar para predecir o estimar la **respuesta media** $\mu_{Y|x_0}$ en $x = x_0$, donde x_0 no necesariamente es uno de los valores preestablecidos, o cuando $x = x_0$, se podría emplear para pronosticar un solo valor y_0 de la variable Y_0. Se esperaría que el error de predicción fuera mayor para el caso de un solo valor pronosticado que para aquel en que se predice una media. Entonces, esto afectaría la anchura de los intervalos para los valores que se predicen.

Suponga que el experimentador desea construir un intervalo de confianza para $\mu_{Y|x_0}$. En tal caso debe usar el estimador puntual $\hat{Y}_0 = B_0$ y B_1x_0 para estimar $\mu_{Y|x_0} = \beta_0 + \beta_1x$. Se puede demostrar que la distribución muestral de $\hat{Y}_0$ es normal con media

$$\mu_{Y|x_0} = E(\hat{Y}_0) = E(B_0 + B_1x_0) = \beta_0 + \beta_1x_0 = \mu_{Y|x_0}$$

y varianza

$$\sigma^2_{\hat{Y}_0} = \sigma^2_{B_0+B_1x_0} = \sigma^2_{\bar{Y}+B_1(x_0-\bar{x})} = \sigma^2\left[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right],$$

esta última surge del hecho de que $\text{Cov}(\bar{Y}_0, B_1) = 0$ (véase el ejercicio de repaso 11.61 de la página 438). Por consiguiente, ahora podemos construir un intervalo de confianza de $100(1-\alpha)\%$ sobre la respuesta media $\mu_{Y|x_0}$ a partir del estadístico

$$T = \frac{\hat{Y}_0 - \mu_{Y|x_0}}{S\sqrt{1/n + (x_0-\bar{x})^2/S_{xx}}},$$

que tiene una distribución t con $n-2$ grados de libertad.

Intervalo de confianza para $\mu_{Y|x_0}$

Un intervalo de confianza de $100(1-\alpha)\%$ para la respuesta media $\mu_{Y|x_0}$ es

$$\hat{y}_0 - t_{\alpha/2}s\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}} < \mu_{Y|x_0} < \hat{y}_0 + t_{\alpha/2}s\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}},$$

$t_{\alpha/2}$ es un valor de la distribución t con $n-2$ grados de libertad.

Predicción del intervalo

Otro tipo de intervalo que con frecuencia se malinterpreta y se confunde con aquel dado para $\mu_{Y|x}$ es el intervalo de la predicción para una respuesta futura observada. En realidad, en muchos casos el intervalo de la predicción es más relevante para el científico o el ingeniero que el intervalo de confianza sobre la media. En el ejemplo del contenido de alquitrán y la temperatura de entrada, mencionado en la sección 11.1, seguramente sería interesante no sólo estimar la media del contenido de alquitrán a una temperatura específica, sino también construir un intervalo que refleje el error en la predicción de una cantidad futura observada del contenido de alquitrán a la temperatura dada.

Para obtener un **intervalo de predicción** para cualquier valor único y_0 de la variable Y_0 es necesario estimar la varianza de las diferencias entre las ordenadas $\hat{y}_0$, obtenidas de las rectas de regresión calculadas en el muestreo repetido cuando $x = x_0$, y la ordenada verdadera correspondiente y_0. Podríamos considerar la diferencia $\hat{y}_0 - y_0$ como un valor de la variable aleatoria $\hat{Y}_0 - Y_0$, cuya distribución muestral se podría demostrar que es normal con media

$$\mu_{\hat{Y}_0-Y_0} = E(\hat{Y}_0 - Y_0) = E[B_0 + B_1x_0 - (\beta_0 + \beta_1x_0 + \epsilon_0)] = 0$$

y varianza

$$\sigma^2_{\hat{Y}_0-Y_0} = \sigma^2_{B_0+B_1x_0-\epsilon_0} = \sigma^2_{\bar{Y}+B_1(x_0-\bar{x})-\epsilon_0} = \sigma^2\left[1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right].$$

Así, un intervalo de predicción de $100(1-\alpha)\%$ para un solo valor pronosticado y_0 se puede construir a partir del estadístico

$$T = \frac{\hat{Y}_0 - Y_0}{S\sqrt{1 + 1/n + (x_0 - \bar{x})^2/S_{xx}}},$$

que tiene una distribución t con $n-2$ grados de libertad.

Intervalo de predicción para y_0

Un intervalo de predicción de $100(1-\alpha)\%$ para una sola respuesta y_0 es dado por

$$\hat{y}_0 - t_{\alpha/2}s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} < y_0 < \hat{y}_0 + t_{\alpha/2}s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}},$$

donde $t_{\alpha/2}$ es un valor de la distribución t con $n-2$ grados de libertad.

Es claro que hay una diferencia entre el concepto de un intervalo de confianza y el del intervalo de predicción antes descrito. La interpretación del intervalo de confianza es idéntica a la que se describió para todos los intervalos de confianza sobre los parámetros de la población estudiados en el libro. De hecho, $\mu_{Y|x_0}$ es un parámetro de la población. Sin embargo, el intervalo de la predicción calculado representa un intervalo que tiene una probabilidad igual a $1-\alpha$ de contener no un parámetro sino un valor futuro de y_0 de la variable aleatoria Y_0.

11.6. elección de un modelo de regresión

Gran parte de lo que se ha presentado hasta ahora acerca de la regresión que involucra una sola variable independiente depende de la suposición de que el modelo elegido es correcto, la suposición de que $\mu_{Y|x}$ se relaciona con x linealmente en los parámetros. Es cierto que no se esperaría que la predicción de la respuesta fuera buena si hubiera diversas variables independientes que no se tomaran en cuenta en el modelo, que afectaran la respuesta y variaran en el sistema. Además, la predicción seguramente sería inadecuada si la estructura verdadera que relaciona $\mu_{Y|x}$ con x fuera extremadamente no lineal en el rango de las variables consideradas.

Es frecuente que se utilice el modelo de regresión lineal simple aun cuando se sepa que el modelo no es lineal o que se desconozca la estructura verdadera. Este método suele ser acertado, en particular cuando el rango de las x es estrecho. De esta manera, el modelo que se utiliza se vuelve una función de aproximación que se espera sea una representación adecuada del panorama verdadero en la región de interés. Sin embargo, hay que señalar el efecto que tendría un modelo inadecuado sobre los resultados presentados hasta este momento. Por ejemplo, si el modelo verdadero, desconocido para el experimentador, es lineal en más de una x, digamos,

$$\mu_{Y|x_1, x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

entonces el estimado $b_1 = S_{xy}/S_{xx}$ de los mínimos cuadrados ordinarios que se calcula considerando tan sólo x_1 en el experimento es, en circunstancias generales, un estimado sesgado del coeficiente β_1, donde el sesgo es una función del coeficiente adicional β_2 (véase el ejercicio de repaso 11.65 en la página 438). Asimismo, el estimado s^2 para σ^2 es sesgado debido a la variable adicional.

11.7. procedimiento del análisis de varianza

Con frecuencia el problema de analizar la calidad de la recta de regresión estimada se maneja por medio del método del **análisis de varianza** (ANOVA), que es un procedimiento mediante el cual la variación total de la variable dependiente se subdivide en componentes significativos, que luego se observan y se tratan en forma sistemática. El análisis de varianza, que se estudia en el capítulo 13, es un recurso poderoso que se emplea en muchas situaciones.

Suponga que tenemos n puntos de datos experimentales en la forma usual (x_i, y_i) y que se estima la recta de regresión. En la sección 11.4 para la estimación de σ^2 se estableció la identidad

$$S_{yy} = b_1 S_{xy} + SCE.$$

Una formulación alternativa y quizá más informativa es la siguiente:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Logramos hacer una partición de la **suma total de los cuadrados corregida de *y*** en dos componentes que deberían proporcionar un significado particular para el experimentador. Esta partición se debería indicar en forma simbólica como

$$STCC = SCR + SCE.$$

El primer componente de la derecha, SCR, se denomina **suma de cuadrados de la regresión** y refleja la cantidad de variación de los valores y que se **explica con el modelo**, que en este caso es la línea recta postulada. El segundo componente es la ya conocida suma de cuadrados del error, que refleja la variación alrededor de la recta de regresión.

Suponga que nos interesa probar la hipótesis

$$H_0\colon\ \beta_1 = 0 \text{ en comparación con } H_1\colon\ \beta_1 \neq 0,$$

donde la hipótesis nula en esencia dice que el modelo es $\mu_{Y|x} = \beta_0$; es decir, la variación en los resultados Y debida a las fluctuaciones de probabilidad o aleatorias que son independientes de los valores de x. Esta condición se refleja en la figura 11.10*b*). En las condiciones de esta hipótesis nula se puede demostrar que SCR/σ^2, y SCE/σ^2 son valores de variables chicuadradas independientes con 1 y $n-2$ grados de libertad, respectivamente y, usando el teorema 7.12, se sigue que $STCC/\sigma^2$ también es un valor de una variable chi cuadrada con $n-1$ grados de libertad. Para probar la hipótesis anterior calculamos

$$f = \frac{SCR\,/\,1}{SCE\,/\,(n-2)} = \frac{SCR}{s^2}$$

y rechazamos H_0 al nivel de significancia α cuando $f > f_\alpha(1, n-2)$.

Por lo general los cálculos se resumen mediante las medias de una **tabla de análisis de varianza**, como se indica en la tabla 11.2. Es costumbre referirse a las distintas sumas de los cuadrados divididos entre sus respectivos grados de libertad como **cuadrados medios**.

Tabla 11.2: Análisis de varianza para la prueba de $\beta_1 = 0$

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	f calculada
Regresión	SCR	1	SCR	$\frac{SCR}{s^2}$
Error	SCE	$n-2$	$s^2 = \frac{SCE}{n-2}$	
Total	$STCC$	$n-1$		

Cuando se rechaza la hipótesis nula, es decir, cuando el estadístico F calculado excede al valor crítico $f_\alpha(1, n-2)$, concluimos que **hay una cantidad significativa de variación en la respuesta justificada por el modelo postulado, que es la función de la línea recta**. Si el estadístico F está en la región de no rechazo, se concluye que los datos no reflejan evidencia suficiente para apoyar el modelo que se postula.

En la sección 11.5 se presentó un procedimiento donde se usa el estadístico

$$T = \frac{B_1 - \beta_{10}}{S/\sqrt{S_{xx}}}$$

para probar la hipótesis

$$H_0\colon\ \beta_1 = \beta_{10} \text{ contra } H_1\colon\ \beta_1 \neq \beta_{10},$$

donde T sigue la distribución t con $n-2$ grados de libertad. La hipótesis se rechaza si $|t| > t_{\alpha/2}$ para un nivel de significancia α. Es interesante observar que en el caso especial en que probamos

$$H_0\colon\ \beta_1 = 0 \text{ en comparación con } H_1\colon\ \beta_1 \neq 0,$$

el valor del estadístico T se convierte en

$$t = \frac{b_1}{s/\sqrt{S_{xx}}},$$

y la hipótesis a considerar es idéntica a la que se prueba en la tabla 11.2. En otras palabras, la hipótesis nula establece que la variación en la respuesta se debe tan sólo al azar. El análisis de varianza utiliza la distribución F en vez de la distribución t. Para la alternativa bilateral ambos enfoques son idénticos. Esto se observa si se escribe

$$t^2 = \frac{b_1^2 S_{xx}}{s^2} = \frac{b_1 S_{xy}}{s^2} = \frac{SCR}{s^2},$$

que da como resultado un valor idéntico al valor f utilizado en el análisis de varianza. La relación fundamental entre la distribución t con v grados de libertad y la distribución F con 1 y v grados de libertad es

$$t^2 = f(1, v).$$

Desde luego, la prueba t permite probar en comparación con una alternativa unilateral, en tanto que la prueba F está restringida a una prueba en comparación con una alternativa bilateral.

11.8. Prueba de la linealidad de la regresión datos con observaciones repetidas

En ciertos tipos de situaciones experimentales el investigador tiene la capacidad de efectuar observaciones repetidas de la respuesta para cada valor de x. Aunque no es necesario tener dichas repeticiones para estimar β_0 y β_1, las repeticiones permiten al experimentador obtener información cuantitativa acerca de lo apropiado que resulta el modelo. De hecho, si se generan observaciones repetidas, el investigador puede efectuar una prueba de significancia para determinar si el modelo es o no adecuado.

```
The regression equation is COD = 3.83 + 0.904 Per_Red
Predictor      Coef   SE Coef       T       P
 Constant     3.830     1.768    2.17   0.038
  Per_Red  0.90364   0.05012   18.03   0.000
S = 3.22954    R-Sq = 91.3%    R-Sq(adj ) = 91.0%
              Analysis of Variance
Source            DF       SS       MS        F       P
Regression         1   3390.6   3390.6   325.08   0.000
Residual  Error  31    323.3     10.4
Total             32   3713.9
```

Obs	Per_Red	COD	Fit	SE Fit	Residual	St Resid
1	3.0	5.000	6.541	1.627	-1.541	-0.55
2	36.0	34.000	36.361	0.576	-2.361	-0.74
3	7.0	11.000	10.155	1.440	0.845	0.29
4	37.0	36.000	37.264	0.590	-1.264	-0.40
5	11.0	21.000	13.770	1.258	7.230	2.43
6	38.0	38.000	38.168	0.607	-0.168	-0.05
7	15.0	16.000	17.384	1.082	-1.384	-0.45
8	39.0	37.000	39.072	0.627	-2.072	-0.65
9	18.0	16.000	20.095	0.957	-4.095	-1.33
10	39.0	36.000	39.072	0.627	-3.072	-0.97
11	27.0	28.000	28.228	0.649	-0.228	-0.07
12	39.0	45.000	39.072	0.627	5.928	1.87
13	29.0	27.000	30.035	0.605	-3.035	-0.96
14	40.0	39.000	39.975	0.651	-0.975	-0.31
15	30.0	25.000	30.939	0.588	-5.939	-1.87
16	41.0	41.000	40.879	0.678	0.121	0.04
17	30.0	35.000	30.939	0.588	4.061	1.28
18	42.0	40.000	41.783	0.707	-1.783	-0.57
19	31.0	30.000	31.843	0.575	-1.843	-0.58
20	42.0	44.000	41.783	0.707	2.217	0.70
21	31.0	40.000	31.843	0.575	8.157	2.57
22	43.0	37.000	42.686	0.738	-5.686	-1.81
23	32.0	32.000	32.746	0.567	-0.746	-0.23
24	44.0	44.000	43.590	0.772	0.410	0.13
25	33.0	34.000	33.650	0.563	0.350	0.11
26	45.0	46.000	44.494	0.807	1.506	0.48
27	33.0	32.000	33.650	0.563	-1.650	-0.52
28	46.0	46.000	45.397	0.843	0.603	0.19
29	34.0	34.000	34.554	0.563	-0.554	-0.17
30	47.0	49.000	46.301	0.881	2.699	0.87
31	36.0	37.000	36.361	0.576	0.639	0.20
32	50.0	51.000	49.012	1.002	1.988	0.65
33	36.0	38.000	36.361	0.576	1.639	0.52

Figura 11.14: Salida de resultados de *MINITAB* de la regresión lineal simple para los datos de reducción de la demanda de oxígeno químico; parte I.

Seleccionemos una muestra aleatoria de n observaciones utilizando k valores distintos de x, por ejemplo, $x_1, x_2, \ldots, x_n$, tales que la muestra contenga n_1 valores observados de la variable aleatoria Y_1 correspondientes a los valores x_1, con n_2 valores observados de Y_2 correspondientes a $x_2, \ldots, n_k$ valores observados de Y_k correspondientes a x_k. Necesariamente, $n = \sum_{i=1}^{k} n_i$.

Obs	Fit	SE Fit	95% CI	95% PI
1	6.541	1.627	(3.223, 9.858)	(-0.834, 13.916)
2	36.361	0.576	(35.185, 37.537)	(29.670, 43.052)
3	10.155	1.440	(7.218, 13.092)	(2.943, 17.367)
4	37.264	0.590	(36.062, 38.467)	(30.569, 43.960)
5	13.770	1.258	(11.204, 16.335)	(6.701, 20.838)
6	38.168	0.607	(36.931, 39.405)	(31.466, 44.870)
7	17.384	1.082	(15.177, 19.592)	(10.438, 24.331)
8	39.072	0.627	(37.793, 40.351)	(32.362, 45.781)
9	20.095	0.957	(18.143, 22.047)	(13.225, 26.965)
10	39.072	0.627	(37.793, 40.351)	(32.362, 45.781)
11	28.228	0.649	(26.905, 29.551)	(21.510, 34.946)
12	39.072	0.627	(37.793, 40.351)	(32.362, 45.781)
13	30.035	0.605	(28.802, 31.269)	(23.334, 36.737)
14	39.975	0.651	(38.648, 41.303)	(33.256, 46.694)
15	30.939	0.588	(29.739, 32.139)	(24.244, 37.634)
16	40.879	0.678	(39.497, 42.261)	(34.149, 47.609)
17	30.939	0.588	(29.739, 32.139)	(24.244, 37.634)
18	41.783	0.707	(40.341, 43.224)	(35.040, 48.525)
19	31.843	0.575	(30.669, 33.016)	(25.152, 38.533)
20	41.783	0.707	(40.341, 43.224)	(35.040, 48.525)
21	31.843	0.575	(30.669, 33.016)	(25.152, 38.533)
22	42.686	0.738	(41.181, 44.192)	(35.930, 49.443)
23	32.746	0.567	(31.590, 33.902)	(26.059, 39.434)
24	43.590	0.772	(42.016, 45.164)	(36.818, 50.362)
25	33.650	0.563	(32.502, 34.797)	(26.964, 40.336)
26	44.494	0.807	(42.848, 46.139)	(37.704, 51.283)
27	33.650	0.563	(32.502, 34.797)	(26.964, 40.336)
28	45.397	0.843	(43.677, 47.117)	(38.590, 52.205)
29	34.554	0.563	(33.406, 35.701)	(27.868, 41.239)
30	46.301	0.881	(44.503, 48.099)	(39.473, 53.128)
31	36.361	0.576	(35.185, 37.537)	(29.670, 43.052)
32	49.012	1.002	(46.969, 51.055)	(42.115, 55.908)
33	36.361	0.576	(35.185, 37.537)	(29.670, 43.052)

Figura 11.15: Salida de resultados de *MINITAB* de la regresión lineal simple para los datos de reducción de la demanda de oxígeno químico; parte II.

Definimos

$$y_{ij} = \text{el } j\text{-ésimo valor de la variable aleatoria } Y_i,$$

$$y_{i.} = T_{i.} = \sum_{j=1}^{n_i} y_{ij},$$

$$\bar{y}_{i.} = \frac{T_{i.}}{n_i}.$$

Entonces, si se realizaron $n_4 = 3$ mediciones de Y que corresponden a $x = x_4$, estas observaciones se indicarían por medio de y_{41}, y_{42} y y_{43}. Por lo tanto,

$$T_{i.} = y_{41} + y_{42} + y_{43}.$$

El concepto de la falta de ajuste

La suma de cuadrados del error consta de dos partes: la cantidad debida a la variación entre los valores de Y dentro de valores dados de x, y un componente que normalmente

se denomina contribución a la **falta de ajuste**. El primer componente refleja tan sólo la variación aleatoria, o **error experimental puro**, en tanto que el segundo es una medida de la variación sistemática introducida por los términos de orden superior. En nuestro caso éstos son términos de x distintos de la contribución lineal o de primer orden. Observe que al elegir un modelo lineal en esencia asumimos que este segundo componente no existe y que, en consecuencia, la suma de cuadrados del error se debe por completo a errores aleatorios. Si éste fuera el caso, entonces $s^2 = SCE/(n-2)$ es un estimado insesgado de σ^2. Sin embargo, si el modelo no se ajusta a los datos en forma apropiada, entonces la suma de cuadrados del error estará inflada y producirá un estimador sesgado de σ^2. Ya sea que el modelo se ajuste o no a los datos, siempre que se tienen observaciones repetidas es posible obtener un estimador insesgado de σ^2 calculando

$$s_i^2 = \frac{\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2}{n_i - 1}, \quad i = 1, 2, \ldots, k,$$

para cada uno de los k valores distintos de x y, después, agrupando estas varianzas, tenemos

$$s^2 = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{n - k} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2}{n - k}.$$

El numerador de s^2 es una **medida del error experimental puro**. A continuación se presenta un procedimiento de cálculo para separar la suma de los cuadrados del error en los dos componentes que representan el error puro y la falta de ajuste:

Cálculo de la suma de los cuadrados de la falta de ajuste

1. Calcular la suma de los cuadrados del error puro

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2.$$

Esta suma de cuadrados tiene $n-k$ grados de libertad asociados con ella, y el cuadrado medio resultante es el estimador insesgado s^2 de σ^2.

2. Restar la suma de los cuadrados del error puro de la suma de los cuadrados del error, SCE, con lo que se obtiene la suma de los cuadrados debida a la falta de ajuste. Los grados de libertad de la falta de ajuste también se obtienen simplemente restando $(n-2) - (n-k) = k-2$.

Los cálculos necesarios para probar hipótesis en un problema de regresión con mediciones repetidas de la respuesta se pueden resumir como se muestra en la tabla 11.3.

Las figuras 11.16 y 11.17 ilustran los puntos muestrales para las situaciones del "modelo correcto" y del "modelo incorrecto". En la figura 11.16, donde $\mu_{Y|x}$ cae sobre una línea recta, no hay falta de ajuste cuando se asume un modelo lineal, por lo que la variación muestral alrededor de la recta de regresión es un error puro que resulta de la variación que ocurre entre observaciones repetidas. En la figura 11.17, donde es evidente que $\mu_{Y|x}$ no cae sobre una línea recta, la responsable de la mayor parte de la variación alrededor de la recta de regresión, además del error puro, es la falta de ajuste que resulta de seleccionar por error un modelo lineal.

Tabla 11.3: Análisis de varianza para la prueba de linealidad de la regresión

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	f calculada
Regresión	SCR	1	SCR	$\frac{SCR}{s^2}$
Error	SCE	$n-2$		
Falta de ajuste	SCE-SCE (puro)	$k-2$	$\frac{SCE\text{-}SCE\ (\text{puro})}{k-2}$	$\frac{SCE\text{-}SCE\ (\text{puro})}{s^2(k-2)}$
Error puro	SCE (puro)	$n-k$	$s^2 = \frac{SCE\ (\text{puro})}{n-k}$	
Total	$STCC$	$n-1$		

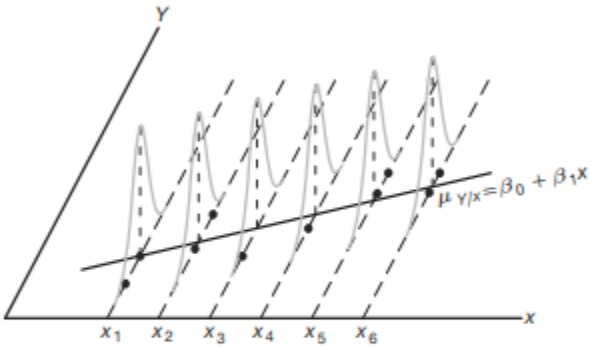


Figura 11.16: Modelo lineal correcto con componente sin falta de ajuste.

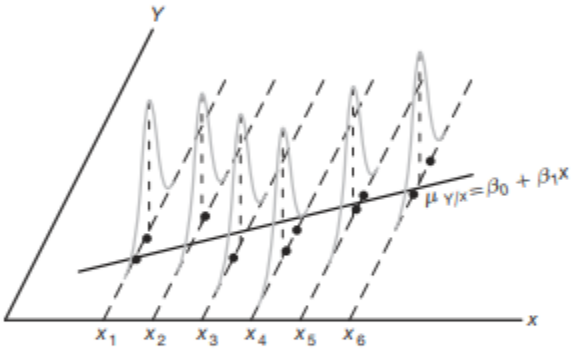


Figura 11.17: Modelo lineal incorrecto con componente de falta de ajuste.

¿Por qué es importante detectar la falta de ajuste?

El concepto de falta de ajuste es muy importante en las aplicaciones del análisis de regresión. De hecho, la necesidad de construir o diseñar un experimento que tome en cuenta la falta de ajuste se vuelve más crítica a medida que el problema y el mecanismo subyacente implicados se vuelven más complicados. Es cierto que no siempre se puede tener la certeza de que la estructura que se postula, en este caso el modelo de regresión lineal, sea una representación correcta o incluso adecuada. El ejemplo siguiente muestra la manera en que se parte la suma de cuadrados del error en los dos componentes que representan el error puro y la falta de ajuste. Lo adecuado del modelo se prueba al nivel de significancia α, comparando el cuadrado medio de la falta de ajuste dividido entre s^2 con $f_\alpha(k-2, n-k)$.

11.9. Gráficas y transformaciones de datos

En este capítulo se estudia la construcción de modelos de regresión en los que hay una variable independiente o regresora. Además, se supone que durante la construcción del modelo tanto x como y entran en el modelo en *forma lineal*. Con frecuencia es aconsejable trabajar con un modelo alternativo en el que x o y (o ambas) intervengan en una forma no lineal. Se podría recomendar una **transformación** de los datos debido a consideraciones teóricas inherentes al estudio científico, o bien, una simple graficación de los datos podría sugerir la necesidad de *reexpresar* las variables en el modelo. La necesidad de llevar a cabo una transformación es muy fácil de diagnosticar en el caso de la regresión lineal simple, ya que las gráficas en dos dimensiones brindan un panorama verdadero de la manera en que las variables se comportan en el modelo.

Un modelo en el que x o y se transforman no debería considerarse como un *modelo de regresión no lineal*. Por lo general denominamos a un modelo de regresión como lineal cuando es **lineal en los parámetros**. En otras palabras, suponga que el aspecto de los datos u otra información científica sugiere que debe hacerse la **regresión de y^* en comparación con la de x^***, donde cada una de ellas es una transformación de las variables naturales x y y. Entonces, el modelo de la forma

$$y_i^* = \beta_0 + \beta_1 x_i^* + \epsilon_i$$

es lineal porque lo es en los parámetros β_0 y β_1. El material que se estudió en las secciones 11.2 a 11.9 permanece sin cambio, donde y_i^* y x_i^* reemplazan a y_i y x_i. Un ejemplo sencillo y útil es el modelo log-log:

$$\log y_i = \beta_0 + \beta_1 \log x_i + \epsilon_i.$$

Aunque este modelo es no lineal en x y y, sí lo es en los parámetros y por ello recibe el tratamiento de un modelo lineal. Por otro lado, un ejemplo de modelo verdaderamente no lineal es:

$$y_i = \beta_0 + \beta_1 x^{\beta_2} + \epsilon_i,$$

donde se debe estimar el parámetro β_2 así como β_0 y β_1. El modelo es no lineal en β_2.

Las transformaciones susceptibles de mejorar el ajuste y la capacidad de predicción de un modelo son muy numerosas. Para un análisis completo de las transformaciones el lector podría consultar a Myers (1990, véase la bibliografía). Decidimos incluir aquí algunas de ellas y mostrar la apariencia de las gráficas que sirven como herramientas diagnósticas. Considere la tabla 11.6, donde se presentan varias funciones que describen relaciones entre y y x que pueden producir una *regresión lineal* por medio de la transformación indicada. Además, en aras de que el análisis sea más exhaustivo, se presentan al lector las variables dependiente e independiente que se utilizan en la *regresión lineal simple* resultante. La figura 11.19 ilustra las funciones que se listan en la tabla 11.6, las cuales sirven como guía para el analista en la elección de una transformación a partir de la observación de la gráfica de y contra x.

Tabla 11.6: Algunas transformaciones útiles para linealizar

Forma funcional que relaciona y con x	**Transformación propia**	**Forma de la regresión lineal simple**
Exponencial: $y = \beta_0 e^{\beta_1 x}$	$y^* = \ln y$	Hacer la regresión de y^* contra x
Potencia: $y = \beta_0 x^{\beta_1}$	$y^* = \log y$; $x^* = \log x$	Hacer la regresión de y^* contra x^*
Recíproca: $y = \beta_0 + \beta_1 \left(\frac{1}{x}\right)$	$x^* = \frac{1}{x}$	Hacer la regresión de y contra x^*
Hiperbólica: $y = \frac{x}{\beta_0 + \beta_1 x}$	$y^* = \frac{1}{y}$; $x^* = \frac{1}{x}$	Hacer la regresión de y^* contra x^*

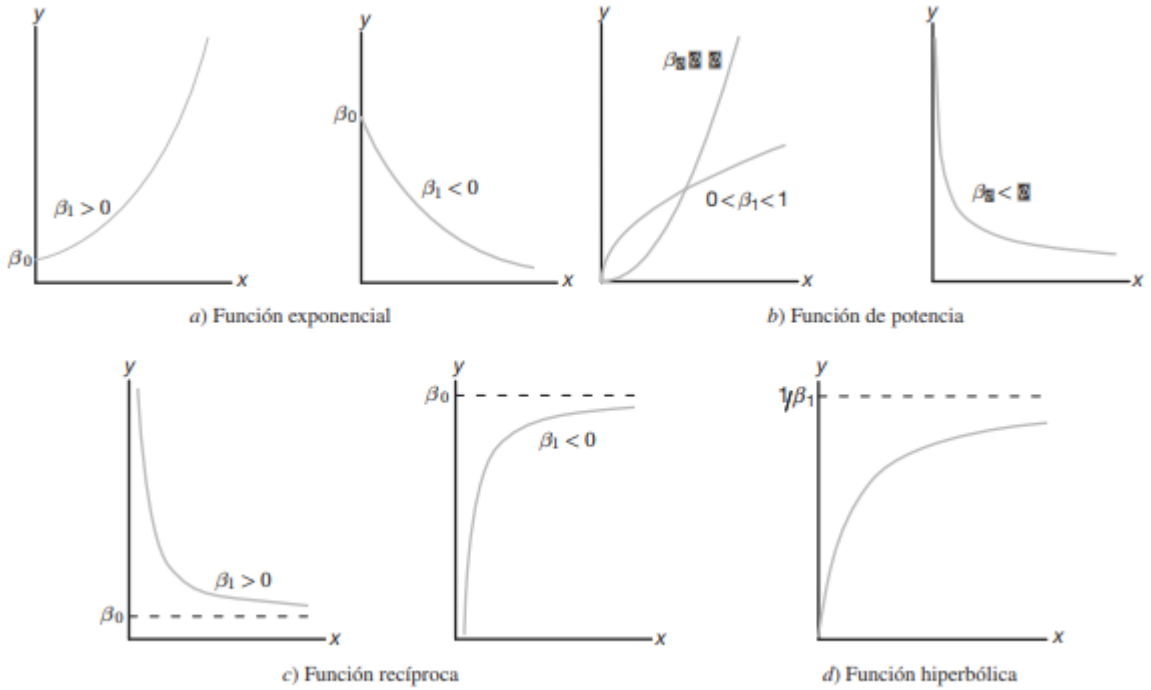


Figura 11.19: Diagramas que ilustran las funciones listadas en la tabla 11.6.

11.10. estudios de regresión lineal simple

En la fabricación de productos comerciales de madera es importante estimar la relación que hay entre la densidad de un producto de madera y su rigidez. Se está considerando un tipo relativamente nuevo de aglomerado que se puede formar con mucha mayor facilidad que el producto comercial ya aceptado. Es necesario saber a qué densidad su rigidez es comparable con la del producto comercial bien conocido y documentado. Terrance E. Conners realizó un estudio titulado *Investigation of Certain Mechanical Properties of a Wood-Foam Composite* (Tesis para el doctorado, Departamento de Bosques y Vida Silvestre, University of Massachusetts). Se produjeron 30 tableros de aglomerado con densidades que variaban aproximadamente de 8 a 26 libras por pie cúbico y se midió su rigidez en libras por pulgada cuadrada. En la tabla 11.8 se presentan los datos.

Es necesario que el analista de datos se concentre en un ajuste apropiado para los datos y que utilice los métodos de inferencia que se estudian en este capítulo. Tal vez lo más apropiado sea una prueba de hipótesis sobre la pendiente de la regresión, así como

la estimación de los intervalos de confianza o de predicción. Se comenzará presentando un simple diagrama de dispersión de los datos brutos con una regresión lineal simple sobrepuesta. En la figura 11.23 se observa dicha gráfica.

El ajuste de regresión lineal simple a los datos produce el modelo ajustado

$$\hat{y} = -25{,}433.739 + 3884.976x \quad (R^2 = 0.7975),$$

Tabla 11.8: Densidad y rigidez de 30 tableros de aglomerado

Densidad, x	**Rigidez, y**	**Densidad, x**	**Rigidez, y**
9.50	14,814.00	8.40	17,502.00
9.80	14,007.00	11.00	19,443.00
8.30	7573.00	9.90	14,191.00
8.60	9714.00	6.40	8076.00
7.00	5304.00	8.20	10,728.00
17.40	43,243.00	15.00	25,319.00
15.20	28,028.00	16.40	41,792.00
16.70	49,499.00	15.40	25,312.00
15.00	26,222.00	14.50	22,148.00
14.80	26,751.00	13.60	18,036.00
25.60	96,305.00	23.40	104,170.00
24.40	72,594.00	23.30	49,512.00
19.50	32,207.00	21.20	48,218.00
22.80	70,453.00	21.70	47,661.00
19.80	38,138.00	21.30	53,045.00

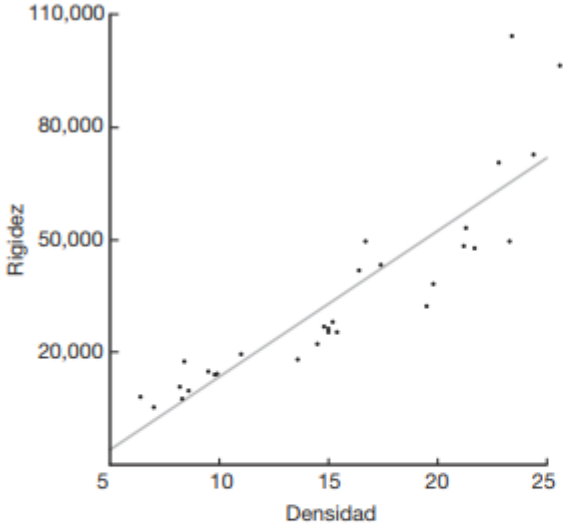


Figura 11.23: Diagrama de dispersión de los datos de densidad de la madera.

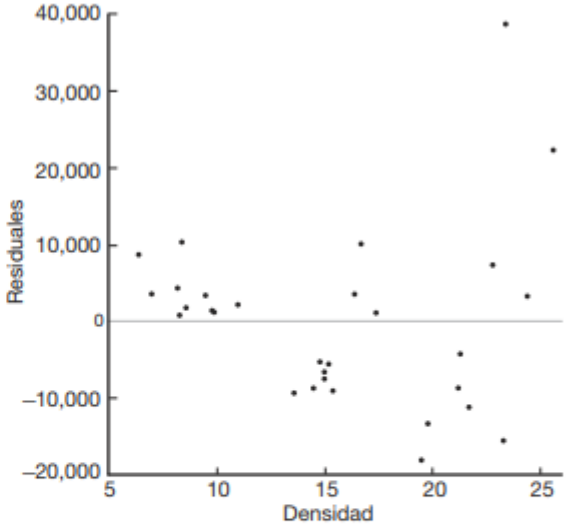


Figura 11.24: Gráfica de los residuales para los datos de densidad de la madera.

y se calcularon los residuales. En la figura 11.24 se presentan los residuales graficados contra las mediciones de la densidad. Difícilmente se trata de un conjunto de residuales ideal o satisfactorio, pues no muestran una distribución aleatoria alrededor del valor de cero. En realidad, los agrupamientos de valores positivos y negativos sugerirían que se debe investigar una tendencia curvilínea en los datos.

Para darnos una idea respecto a la suposición de error normal se dibujó una gráfica de probabilidad normal de los residuales. Es el tipo de gráfica que estudiamos en la sección 8.8, donde el eje horizontal representa la función de distribución normal empírica en una escala que produce una gráfica con línea recta cuando se grafica contra los residuales. En la figura 11.25 se presenta la gráfica de probabilidad normal de los residuales. Esta gráfica no refleja la apariencia de recta que a uno le gustaría ver, lo cual es otro síntoma de una selección errónea, quizá sobresimplificada, de un modelo de regresión.

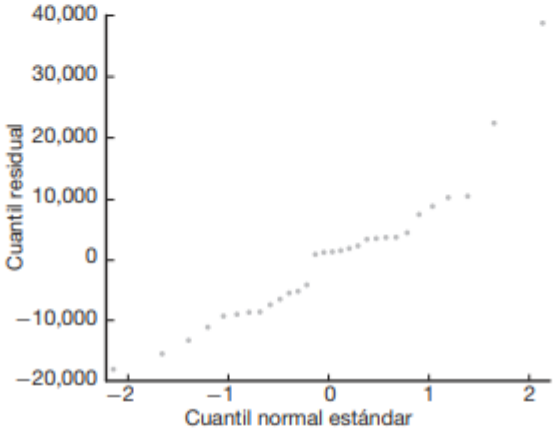


Figura 11.25: Gráfica de probabilidad normal de los residuales para los datos de densidad de la madera.

Los dos tipos de gráficas de residuales y, de hecho, el propio diagrama de dispersión, sugieren que sería adecuado un modelo algo más complicado. Una posibilidad es usar un modelo con transformación de logaritmos naturales. En otras palabras, hay que elegir hacer la regresión de ln y contra x. Esto produce la regresión

$$\widehat{\ln y} = 8.257 + 0.125x \quad (R^2 = 0.9016).$$

Para darse una idea de si el modelo transformado es más apropiado considere las figuras 11.26 y 11.27, que muestran las gráficas de los residuales de la rigidez [es decir, y_i - antilog $(\widehat{\ln y})$]] en comparación con las de la densidad. La figura 11.26 parece más cercana a un patrón aleatorio alrededor del cero, en tanto que la figura 11.27 con seguridad se acerca más a una línea recta. Esto, además de un valor de R^2 más elevado, sugeriría que el modelo transformado es más apropiado.

11.11. Correlación

Hasta este momento se ha supuesto que la variable regresora independiente x es una variable científica o física en lugar de una variable aleatoria. De hecho, en este contexto es frecuente que x se denomine **variable matemática**, la cual, en el proceso de muestreo, se mide con un error despreciable. En muchas aplicaciones de las técnicas de regresión es más realista suponer que tanto X como Y son variables aleatorias y que las mediciones $\{(x_i, y_i); i = 1, 2, \ldots, n\}$ son observaciones de una población que tiene la función de

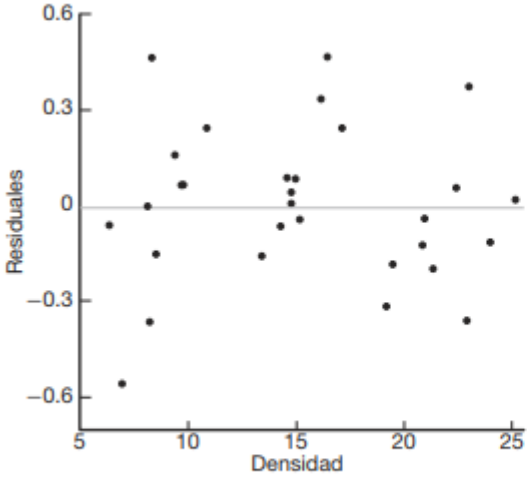


Figura 11.26: Gráfica de residuales donde se utiliza una transformación logarítmica para los datos de densidad de la madera.

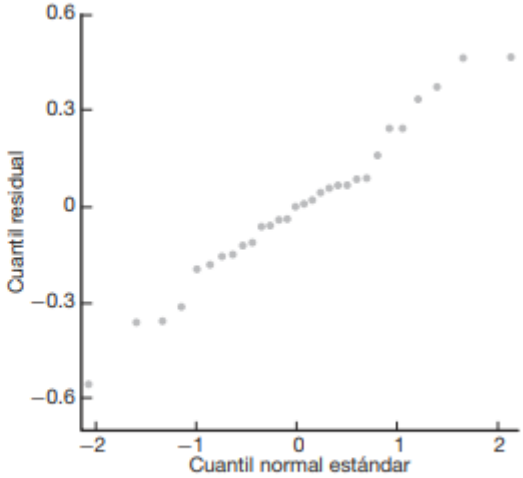


Figura 11.27: Gráfica de probabilidad normal de residuales en la cual se utiliza una transformación logarítmica para los datos de densidad de la madera.

densidad conjunta $f(x, y)$. Debemos considerar el problema de medir la relación entre las dos variables X y Y. Por ejemplo, si X y Y representaran la longitud y la circunferencia de una clase particular de hueso en el cuerpo de un adulto, podríamos realizar un estudio antropológico para determinar si los valores grandes de X se asocian con valores grandes de Y, y viceversa.

Por otro lado, si X representa la antigüedad de un automóvil usado y Y representa su precio de lista al menudeo, se esperaría que los valores grandes de X correspondan a valores pequeños de Y y que los valores pequeños de X correspondan a valores grandes de Y. El **análisis de correlación** intenta medir la fuerza de tales relaciones entre dos variables por medio de un solo número denominado **coeficiente de correlación**.

En teoría, con frecuencia se supone que la distribución condicional $f(y|x)$ de Y, para valores fijos de X, es normal con media $\mu_{Y|x} = \beta_0 + \beta_1 x$ y varianza $\sigma^2_{Y|x} = \sigma^2$, y que, de igual manera, X se distribuye de forma normal con media μ y varianza σ^2_x. Entonces, la densidad conjunta de X y Y es

$$f(x, y) = n(y|x; \beta_0 + \beta_1 x, \sigma)n(x; \mu_X, \sigma_X)$$

$$= \frac{1}{2\pi\sigma_x\sigma} \exp\left\{-\frac{1}{2}\left[\left(\frac{y-\beta_0-\beta_1 x}{\sigma}\right)^2 + \left(\frac{x-\mu_X}{\sigma_X}\right)^2\right]\right\},$$

para $-\infty < x < \infty$ y $-\infty < y < \infty$.

Escribamos la variable aleatoria Y en la forma

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

donde ahora X es una variable aleatoria independiente del error aleatorio ϵ. Como la media del error aleatorio ϵ es cero, se deduce que

$$\mu_Y = \beta_0 + \beta_1\mu_X \quad \text{y} \quad \sigma^2_Y = \sigma^2 + \beta_1^2\sigma_X^2 .$$

Al sustituir para α y σ^2 en la expresión anterior para $f(x, y)$, se obtiene la **distribución normal bivariada**

$$f(x, y) = \frac{1}{2\pi\sigma_X \sigma_Y \sqrt{1-\rho^2}}$$
$$\times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\},$$

para $-\infty < x < \infty$ y $-\infty < y < \infty$, donde

$$\rho^2 = 1 - \frac{\sigma^2}{\sigma_Y^2} = \beta_1^2 \frac{\sigma_X^2}{\sigma_Y^2}.$$

La constante ρ (ro) se denomina **coeficiente de correlación de la población** y desempeña un papel importante en muchos problemas de análisis de datos bivariados. Es importante que el lector entienda la interpretación física de este coeficiente de correlación, así como la diferencia entre correlación y regresión. El término *regresión* aún tiene algún significado aquí. De hecho, la línea recta dada por $\mu_{Y|x} = \beta_0 + \beta_1 x$ se sigue llamando recta de regresión, igual que antes, y los estimadores de β_0 y β_1 son idénticos a los que se presentaron en la sección 11.3. El valor de ρ es 0 cuando $\beta_1 = 0$, que resulta cuando en esencia no existe regresión lineal; es decir, cuando la recta de regresión es horizontal y cualquier conocimiento de X es inútil para predecir Y. Como $\sigma_Y^2 \geq \sigma^2$, se debe tener $\rho^2 \leq 1$ y, por lo tanto, $-1 \leq \rho \leq 1$. Los valores de ρ ± 1 sólo ocurren cuando $\sigma^2 = 0$, en cuyo caso se tiene una relación lineal perfecta entre las dos variables. Así, un valor de ρ igual a $+1$ implica una relación lineal perfecta con pendiente positiva, en tanto que un valor de ρ igual a -1 resulta de una relación lineal perfecta con pendiente negativa. Entonces, se podría decir que los estimadores muestrales de ρ con magnitud cercana a la unidad implican una buena correlación o **asociación lineal** entre X y Y, mientras que valores cercanos a cero indican poca o ninguna correlación.

Para obtener un estimador muestral de ρ recordemos que en la sección 11.4 aprendimos que la suma de los cuadrados del error es

$$SCE = S_{yy} - b_1 S_{xy}.$$

Al dividir ambos lados de esta ecuación entre S_{yy} y reemplazar S_{xy} con $b_1 S_{xx}$, se obtiene la relación

$$b_1^2 \frac{S_{xx}}{S_{yy}} = 1 - \frac{SCE}{S_{yy}}.$$

El valor de $b_1^2 S_{xx} / S_{yy}$ es igual a cero cuando $b_1 = 0$, lo que ocurrirá cuando los puntos muestrales no tengan relación lineal. Como $S_{yy} \geq SCE$, se concluye que $b_1^2 S_{xx} / S_{xy}$ debe estar entre 0 y 1. En consecuencia, $b_1\sqrt{S_{xx} / S_{yy}}$ debe variar entre -1 y $+1$, y los valores negativos corresponden a rectas con pendientes negativas, mientras que los valores positivos corresponden a rectas con pendientes positivas. Un valor de -1 o $+1$ sucederá cuando $SCE = 0$, pero éste es el caso en el que todos los puntos muestrales caen sobre una línea recta. Por lo tanto, una relación lineal perfecta se da en los datos muestrales cuando $b_1\sqrt{S_{xx} / S_{yy}} = \pm 1$. Es claro que la cantidad $b_1\sqrt{S_{xx} / S_{yy}}$, la cual se designará de aquí en adelante como r, se puede usar como un estimado del coeficiente de correlación ρ de la población. Se acostumbra hacer referencia al estimado r como **coeficiente de correlación producto-momento de Pearson**, o sólo como **coeficiente de correlación muestral**.

Coeficiente de correlación

La medida ρ de la asociación lineal entre dos variables X y Y se estima por medio del **coeficiente de correlación muestral** r, donde

$$r = b_1\sqrt{\frac{S_{xx}}{S_{yy}}} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}.$$

Hay que tener cuidado en la interpretación de valores de r entre -1 y $+1$. Por ejemplo, valores de r iguales a 0.3 y 0.6 significan sólo que hay dos correlaciones positivas, una un poco más fuerte que la otra. Sería un error concluir que $r = 0.6$ indica una relación lineal dos veces mejor que la del valor $r = 0.3$. Por otro lado, si escribimos

$$r^2 = \frac{S_{xy}^2}{S_{xx}\, S_{yy}} = \frac{SCR}{S_{yy}},$$

entonces r^2, que por lo general se denomina **coeficiente muestral de determinación**, representa la proporción de la variación de S_{yy} explicada por la regresión de Y sobre x, a saber, la SCR. Es decir, r^2 expresa la proporción de la variación total de los valores de la variable Y que son ocasionados o explicados por una relación lineal con los valores de la variable aleatoria X. Así, una correlación de 0.6 significa que 0.36, o 36%, de la variación total de los valores de Y en la muestra se explica mediante la relación lineal con los valores de X.

12. REGRESIÓN LINEAL MÚLTIPLE

12.1.Introducción

En la mayoría de los problemas de investigación en los que se aplica el análisis de regresión se necesita más de una variable independiente para el modelo de regresión. La complejidad de la mayoría de mecanismos científicos es tal que, con el fin de predecir una respuesta importante, se requiere un **modelo de regresión múltiple**. Cuando un modelo es lineal en los coeficientes se denomina **modelo de regresión lineal múltiple**. Para el caso de k variables independientes, el modelo que da $x_1, x_2, \ldots, x_k$, la media de $Y|x_1, x_2, \ldots, x_k$ es el modelo de regresión lineal múltiple

$$\mu_{Y|x_1, x_2, \ldots, x_k} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k,$$

y la respuesta estimada se obtiene a partir de la ecuación de regresión muestral

$$\hat{y} = b_0 + b_1 x_1 + \cdots + b_k x_k,$$

donde cada coeficiente de regresión β_i se estima por medio de b_i, a partir de los datos muestrales, usando el método de los mínimos cuadrados. Como ocurre en el caso de una sola variable independiente, a menudo el modelo de regresión lineal múltiple es una representación adecuada de una estructura más complicada dentro de ciertos rangos de las variables independientes.

También se pueden aplicar técnicas similares de mínimos cuadrados para estimar los coeficientes cuando el modelo lineal incluye, por ejemplo, potencias y productos de las variables independientes. Un ejemplo de esto se presentaría cuando $k = 1$, en cuyo caso el experimentador podría pensar que las medias $\mu_{Y|x}$ no caen sobre una línea recta, sino que se describen de manera más adecuada mediante el **modelo de regresión polinomial**

$$\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_r x^r,$$

y la respuesta estimada se obtiene de la ecuación de regresión polinomial

$$\hat{y} = b_0 + b_1 x + b_2 x^2 + \cdots + b_r x^r.$$

En ocasiones se genera confusión al decir que un modelo polinomial es un modelo lineal. Sin embargo, los estadísticos normalmente se refieren a un modelo lineal como aquel en el que los parámetros ocurren en forma lineal, independientemente de cómo las variables independientes entran en el modelo. Un ejemplo de modelo no lineal es la **relación exponencial**

$$\mu_{Y|x} = \alpha\beta^x,$$

que se estima mediante la ecuación de regresión

$$\hat{y} = ab^x.$$

En ciencias e ingeniería hay muchos fenómenos cuya naturaleza no es inherentemente lineal y, cuando se conoce su verdadera estructura, no hay duda de que habría que intentar ajustar el modelo real. Existe mucha literatura acerca de la estimación de modelos no lineales por medio de mínimos cuadrados. Los modelos no lineales que se analizan en este capítulo se relacionan con condiciones no ideales, en las cuales el analista está seguro de que la respuesta y, por lo tanto, el error de respuesta del modelo no se distribuyen normalmente sino que, más bien, tienen una distribución binomial o de Poisson. Estas situaciones ocurren a menudo en la práctica.

El estudiante que busque profundizar en la explicación de la regresión no lineal debe consultar la obra de Myers *Classical and Modern Regression with Applications* (1990; véase la bibliografía).

12.2.estimación de coeficientes

En esta sección se calculan los estimadores de mínimos cuadrados de los parámetros β_0, β_1,..., β_k mediante el ajuste del modelo de regresión lineal múltiple

$$\mu_{Y|x_1,x_2,\ldots,x_k} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

a los puntos de los datos

$$\{(x_{1i}, x_{2i}, \ldots, x_{ki}, y_i); \quad i = 1, 2, \ldots, n \text{ y } n > k\},$$

donde y_i es la respuesta observada a los valores x_{1i}, x_{2i},..., x_{ki} de las k variables independientes x_1, x_2,..., x_k. Se supone que cada observación (x_{1i}, x_{2i},..., x_{ki}, y_i) satisface la siguiente ecuación:

Modelo de regresión lineal múltiple

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i$$

o bien,

$$y_i = \hat{y}_i + e_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki} + e_i,$$

donde ϵ_i y e_i son el error aleatorio y el residual, respectivamente, asociados con la respuesta y_i y con el valor ajustado $\hat{y}_i$.

Como en el caso de la regresión lineal simple, se supone que los ϵ_i son independientes y están distribuidos en forma idéntica con media cero y varianza común σ^2.

Si usamos el concepto de mínimos cuadrados para obtener los estimados b_0, b_1,..., b_k, minimizamos la expresión

$$SCE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - \cdots - b_k x_{ki})^2.$$

Si, a su vez, diferenciamos la *SCE* respecto a b_0, b_1,..., b_k e igualamos el resultado a cero, generamos el conjunto de $k + 1$ **ecuaciones normales para la regresión lineal múltiple**.

Ecuaciones normales de estimación para la regresión lineal múltiple

$$
\begin{aligned}
nb_0 + b_1\sum_{i=1}^{n} x_{1i} + b_2\sum_{i=1}^{n} x_{2i} + \cdots + b_k\sum_{i=1}^{n} x_{ki} &= \sum_{i=1}^{n} y_i \\
b_0\sum_{i=1}^{n} x_{1i} + b_1\sum_{i=1}^{n} x_{1i}^2 + b_2\sum_{i=1}^{n} x_{1i}x_{2i} + \cdots + b_k\sum_{i=1}^{n} x_{1i}x_{ki} &= \sum_{i=1}^{n} x_{1i}y_i \\
\vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots \\
b_0\sum_{i=1}^{n} x_{ki} + b_1\sum_{i=1}^{n} x_{ki}x_{1i} + b_2\sum_{i=1}^{n} x_{ki}x_{2i} + \cdots + b_k\sum_{i=1}^{n} x_{ki}^2 &= \sum_{i=1}^{n} x_{ki}y_i
\end{aligned}
$$

Estas ecuaciones se pueden resolver para $b_0, b_1, b_2, \ldots, b_k$ utilizando cualquier método apropiado que permita resolver sistemas de ecuaciones lineales. Casi todos los programas estadísticos de cómputo se pueden utilizar para obtener soluciones numéricas de las ecuaciones anteriores.

Regresión polinomial

Ahora suponga que se desea ajustar la ecuación polinomial

$$\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_r x^r$$

a los n pares de observaciones $\{(x_i, y_i); i = 1, 2, \ldots, n\}$. Cada observación, y_i, satisface la ecuación

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_r x_i^r + \epsilon_i$$

o bien,

$$y_i = \hat{y}_i + e_i = b_0 + b_1 x_i + b_2 x_i^2 + \cdots + b_r x_i^r + e_i,$$

donde r es el grado del polinomio y ϵ_i y e_i son, de nuevo, el error aleatorio y el residual asociados con la respuesta y_i y con el valor ajustado $\hat{y}_i$, respectivamente. Aquí el número de pares, n, debe ser al menos $r + 1$, que es el número de parámetros por estimar.

Observe que el modelo polinomial se puede considerar un caso especial del modelo de regresión lineal múltiple más general, donde establecemos $x_1 = x, x_2 = x^2, \ldots, x_r = x^r$. Las ecuaciones normales adoptan la misma forma que las que aparecen en la página 445. Luego se resuelven para $b_0, b_1, b_2, \ldots, b_r$.

12.3. modelo de regresión lineal con el uso de matrices

Al ajustar un modelo de regresión lineal múltiple, en particular cuando contiene más de dos variables, tener conocimientos sobre la teoría de matrices facilita considerablemente el manejo de las matemáticas. Suponga que el experimentador tiene k variables

independientes $x_1, x_2, \ldots, x_k$ y n observaciones $y_1, y_2, \ldots, y_n$, cada una de las cuales se puede expresar con la ecuación

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i.$$

Este modelo representa en esencia a n ecuaciones que describen cómo se generan los valores de la respuesta durante el proceso científico. Si usamos la notación de matrices, podemos escribir la ecuación siguiente

Modelo lineal general

$$\mathbf{y} = \mathbf{X}\beta + \epsilon,$$

donde

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Después, el método de mínimos cuadrados para la estimación de β, que se estudió en la sección 12.2, implica calcular $\mathbf{b}$, para lo cual

$$SCE = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$$

se minimiza. Este proceso de minimización implica resolver para $\mathbf{b}$ en la ecuación

$$\frac{\partial}{\partial \mathbf{b}}(SCE) = \mathbf{0}.$$

Aquí no presentaremos los detalles respecto a cómo se resuelven las ecuaciones anteriores. El resultado se reduce a la solución de $\mathbf{b}$ en

$$(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{y}.$$

Observe la naturaleza de la matriz $\mathbf{X}$. Además del elemento inicial, el i-ésimo renglón representa los valores de x que dan lugar a la respuesta y_i. Si escribimos

$$\mathbf{A} = \mathbf{X}'\mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{2i} & \cdots & \sum_{i=1}^{n} x_{ki} \\ \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{1i}^2 & \sum_{i=1}^{n} x_{1i}x_{2i} & \cdots & \sum_{i=1}^{n} x_{1i}x_{ki} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_{i=1}^{n} x_{ki} & \sum_{i=1}^{n} x_{ki}x_{1i} & \sum_{i=1}^{n} x_{ki}x_{2i} & \cdots & \sum_{i=1}^{n} x_{ki}^2 \end{bmatrix}$$

y

$$\mathbf{g} = \mathbf{X}'\mathbf{y} = \begin{bmatrix} g_0 = \sum_{i=1}^{n} y_i \\ g_1 = \sum_{i=1}^{n} x_{1i}y_i \\ \vdots \\ g_k = \sum_{i=1}^{n} x_{ki}y_i \end{bmatrix}$$

nos permite escribir las ecuaciones normales en la forma de matriz

$$\mathbf{A}\mathbf{b} = \mathbf{g}.$$

Si la matriz **A** es no singular, la solución para los coeficientes de regresión se escribe como

$$\mathbf{b} = \mathbf{A}^{-1}\mathbf{g} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

De esta manera, obtenemos la ecuación de predicción o regresión resolviendo un conjunto de $k + 1$ ecuaciones con un número igual de incógnitas. Esto implica el invertir la matriz $\mathbf{X}'\mathbf{X}$ de orden $k + 1$ por $k + 1$. En la mayoría de libros que tratan sobre determinantes y matrices elementales se explican las técnicas para invertir matrices. Por supuesto, existen muchos paquetes de cómputo veloces para resolver problemas de regresión múltiple, los cuales no sólo proporcionan estimados de los coeficientes de regresión, sino que también ofrecen otra clase de información relevante para hacer inferencias acerca de la ecuación de regresión.

12.4. propiedades de los estimadores de mínimos cuadrados

Las medias y varianzas de los estimadores $b_0, b_1, \ldots, b_k$ se obtienen con facilidad si se hacen ciertas suposiciones sobre los errores aleatorios $\epsilon_1, \epsilon_2, \ldots, \epsilon_k$, que son idénticas a las que se hacen en el caso de la regresión lineal simple. Si suponemos que dichos errores son independientes, cada uno con media igual a cero y varianza σ^2, entonces podemos demostrar que $b_0, b_1, \ldots, b_k$ son, respectivamente, estimadores no sesgados de los coeficientes de regresión $\beta_0, \beta_1, \ldots, \beta_k$. Además, las varianzas de las b se obtienen por medio de los elementos del inverso de la matriz **A**. Observe que los elementos fuera de la diagonal de $\mathbf{A} = \mathbf{X}'\mathbf{X}$ representan sumas de productos de los elementos en las columnas de **X**; mientras que los elementos en la diagonal de **A** son las sumas de los cuadrados de los elementos en las columnas de **X**. La matriz inversa, $\mathbf{A}^{-1}$, aparte del multiplicador σ^2, representa la **matriz de varianza-covarianza** de los coeficientes de regresión estimados. Es decir, los elementos de la matriz $\mathbf{A}^{-1}\sigma^2$ muestran las varianzas de $b_0, b_1, \ldots, b_k$ en la diagonal principal y las covarianzas fuera de la diagonal. Por ejemplo, en un problema de regresión lineal múltiple con $k = 2$ se podría escribir

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} c_{00} & c_{01} & c_{02} \\ c_{10} & c_{11} & c_{12} \\ c_{20} & c_{21} & c_{22} \end{bmatrix}$$

con los elementos debajo de la diagonal principal determinados por la simetría de la matriz. Entonces, se escribe

$$\sigma_{b_i}^2 = c_{ii}\,\sigma^2, \qquad i = 0, 1, 2,$$
$$\sigma_{b_i b_j} = \mathrm{Cov}(b_i, b_j) = c_{ij}\,\sigma^2, \quad i \neq j.$$

Desde luego, los estimados de las varianzas y, por lo tanto, sus errores estándar, se obtienen reemplazando σ^2 con el estimado apropiado, el cual se obtuvo a partir de los datos experimentales. Un estimado no sesgado de σ^2 de nuevo se define en términos de

la suma de cuadrados del error, que se calcula utilizando la fórmula establecida en el teorema 12.1. En el teorema las suposiciones se basan en los ϵ_i descritos con anterioridad.

Teorema 12.1: Para la ecuación de regresión lineal

$$\mathbf{y} = \mathbf{X}\beta + \epsilon,$$

un estimador insesgado de σ^2 es dado por el error o media cuadrática residual

$$s^2 = \frac{SCE}{n-k-1}, \quad \text{donde} \quad SCE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Podemos ver que, para el caso de la regresión lineal simple, el teorema 12.1 representa una generalización del teorema 11.1. La prueba se deja como ejercicio para el lector. Al igual que en el caso de la regresión lineal más simple, el estimado de s^2 es una medida de la variación de los errores de la predicción o residuales. En las secciones 12.10 y 12.11 se presentan otras inferencias importantes relacionadas con la ecuación ajustada de regresión, con base en los valores de los residuales individuales $e_i = y_i - \hat{y}_i$, $i = 1, 2,..., n$.

La suma de cuadrados del error y de la regresión adoptan la misma forma y desempeñan el mismo papel que en el caso de la regresión lineal simple. De hecho, la identidad de la suma de cuadrados

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

se sigue cumpliendo, y se conserva la notación anterior, que es,

$$STCC = SCR + SCE,$$

con

$$STCC = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \text{suma de cuadrados total}$$

y

$$SCR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \text{suma de cuadrados de regresión}$$

Hay k grados de libertad asociados con la SCR, y, como siempre, la $STCC$ tiene $n - 1$ grados de libertad. Por lo tanto, después de restar, la SCE tiene $n - k - 1$ grados de libertad. Así, nuestro estimado de σ^2 de nuevo es dado por la suma de cuadrados del error dividida entre sus grados de libertad. Las tres sumas de cuadrados aparecen en la salida de resultados de la mayoría de los programas de cómputo de regresión múltiple. Observe que la condición $n > k$ en la sección 12.2 garantiza que los grados de libertad de la SCE no sean negativos.

Análisis de varianza en la regresión múltiple

La partición de la suma total de cuadrados en sus componentes, la suma de cuadrados de regresión y del error desempeña un papel importante. Puede efectuarse un **análisis de varianza** que arroje luz sobre la calidad de la ecuación de regresión. Una hipótesis que sirve para determinar si el modelo explica una cantidad significativa de variación, es la siguiente:

$$H_0\colon\ \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0\,.$$

El análisis de varianza implica una prueba F, mediante una tabla, como la siguiente:

Fuente	Suma de cuadrados	Grados de libertad	Cuadrados medios	F
Regresión	SCR	k	$CMR = \frac{SCR}{k}$	$f = \frac{CMR}{CME}$
Error	SCE	$n-(k+1)$	$CME = \frac{SCE}{n-(k+1)}$	
Total	$STCC$	$n-1$		

Se trata de una **prueba de cola superior**. El rechazo de H_0 significa que la **ecuación de regresión difiere de una constante**. Es decir, al menos una variable regresora es importante. En las secciones que siguen se estudia más el uso del análisis de varianza.

Otra utilidad del cuadrado medio del error (o cuadrado medio residual) estriba en su uso para la prueba de hipótesis y la estimación de intervalos de confianza que se estudian en la sección 12.5. Además, el cuadrado medio del error desempeña un papel importante en las situaciones en las que el científico busca el mejor modelo entre un conjunto de ellos que están en competencia. Muchos criterios de construcción de modelos incluyen el estadístico s^2. En la sección 12.11 se presentan criterios para comparar modelos en competencia.

12.5.inferencias en la regresión lineal múltiple

El conocimiento de la distribución de los estimadores del coeficiente individual facilita al experimentador construir intervalos de confianza para los coeficientes y hacer pruebas de hipótesis acerca de ellos. Recuerde que en la sección 12.4 estudiamos que b_j ($j = 0, 1, 2, \ldots, k$) se distribuyen de forma normal con media β_j y varianza $c_{jj}\sigma^2$. De esta manera, se puede utilizar el estadístico

$$t = \frac{b_j - \beta_{j0}}{s\sqrt{c_{jj}}}$$

con $n - k - 1$ grados de libertad para probar hipótesis y construir intervalos de confianza sobre β_j. Por ejemplo, si queremos probar

$$H_0:\ \beta_j = \beta_{j0},$$
$$H_1:\ \beta_j \neq \beta_{j0},$$

se calcula el estadístico t anterior y no se rechaza H_0 si $-t_{\alpha/2} < t < t_{\alpha/2}$, donde $t_{\alpha/2}$ tiene $n - k - 1$ grados de libertad.

Ejemplo 12.5: Para el modelo del ejemplo 12.4 pruebe la hipótesis de que $\beta_2 = -2.5$ en comparación con la alternativa de que $\beta_2 > -2.5$ a un nivel de significancia de 0.05.

***Solución*:**

$$H_0:\ \beta_2 = -2.5,$$
$$H_1:\ \beta_2 > -2.5.$$

Cálculos:

$$t = \frac{b_2 - \beta_{20}}{s\sqrt{c_{22}}} = \frac{-1.8616 + 2.5}{2.073\sqrt{0.0166}} = 2.390,$$
$$P = P(T > 2.390) = 0.04.$$

Decisión: Rechazar H_0 y concluir que $\beta_2 > -2.5$.

Pruebas *t* individuales para la selección de variables

La prueba t que se utiliza con más frecuencia en la regresión múltiple es aquella que prueba la importancia de los coeficientes individuales, es decir, H_0: $\beta_j = 0$ en comparación con la hipótesis alternativa H_1: $\beta_j \neq 0$. Con frecuencia estas pruebas contribuyen a lo que se denomina **selección de variables**, con la cual el analista intenta llegar al modelo más útil, es decir, a la elección de cuál regresor utilizar. Aquí debemos destacar que, si se encuentra que un coeficiente es insignificante, es decir, si **no se rechaza** la hipótesis H_0: $\beta_j = 0$, la conclusión que se obtiene es que la **variable** es insignificante (explica una cantidad insignificante de la variación de y) **en la presencia de los demás regresores del modelo**. Más adelante se profundizará en este punto.

Inferencias sobre la respuesta media y la predicción

Una de las inferencias más útiles que se pueden hacer con respecto a la calidad de la respuesta predicha y_0, correspondiente a los valores $x_{10}, x_{20}, \ldots, x_{k0}$, es el intervalo de confianza sobre la respuesta media $\mu_Y | x_{10}, x_{20}, \ldots, x_{k0}$. Estamos interesados en construir un intervalo de confianza sobre la respuesta media para el conjunto de condiciones determinadas por

$$\mathbf{x}_0' = [1, x_{10}, x_{20}, \ldots, x_{k0}].$$

Se aumentan en 1 las condiciones sobre las x para facilitar la notación de matrices. La normalidad en los ϵ_i producen normalidad en los b_j, y la media y la varianza siguen siendo las mismas, como se indica en la sección 12.4. Así es la covarianza entre b_i y b_j para $i \neq j$. De esta manera,

$$\hat{y} = b_0 + \sum_{j=1}^{k} b_j x_{j0}$$

también se distribuye normalmente y es, de hecho, un estimador no sesgado para la **respuesta media** sobre la que se intenta ligar un intervalo de confianza. La varianza de $\hat{y}_0$, escrita con notación de matriz simplemente como función de σ^2, $(\mathbf{X}'\mathbf{X})^{-1}$, y el vector de condiciones, $\mathbf{x}_0'$ es

$$\sigma^2_{\hat{y}_0} = \sigma^2 \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0.$$

Si esta expresión se extendiera para un caso dado, por ejemplo $k = 2$, ya vimos que explica de manera apropiada la varianza de b_j y la covarianza de b_i y b_j, para $i \neq j$. Después de sustituir σ^2 con s^2, según se plantea en el teorema 12.1, el intervalo de confianza del $100(1-\alpha)\%$ se puede construir sobre $\mu_{Y|x} = x_{10}, x_{20}, \ldots, x_{k0}$ a partir del estadístico

$$T = \frac{\hat{y}_0 - \mu_{Y|x_{10}, x_{20}, \ldots, x_{k0}}}{s\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}},$$

que tiene una distribución t con $n - k - 1$ grados de libertad.

Intervalo de confianza para $\mu_{Y|x_{10}, x_{20}, \ldots, x_{k0}}$

Un intervalo de confianza de $100(1-\alpha)\%$ para la **respuesta media** $\mu_{Y|x_{10}, x_{20}, \ldots, x_{k0}}$ es

$$\hat{y}_0 - t_{\alpha/2}s\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} < \mu_{Y|x_{10}, x_{20}, \ldots, x_{k0}} < \hat{y}_0 + t_{\alpha/2}s\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0},$$

donde $t_{\alpha/2}$ es un valor de la distribución t con $n - k - 1$ grados de libertad.

Es frecuente que a la cantidad $s\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}$ se le denomine **error estándar de la predicción** y aparece en la salida de resultados de muchos paquetes de cómputo para regresión.

12.6. elección de un modelo de ajuste la prueba de hipótesis

En muchas situaciones de regresión los coeficientes individuales revisten importancia para el experimentador. Por ejemplo, en una aplicación de economía, $\beta_1, \beta_2, \ldots$ podrían tener un significado en particular, por lo que el economista tendría un interés especial en los intervalos de confianza y en las pruebas de hipótesis sobre dichos parámetros. Sin embargo, considere una situación de química industrial en la que el modelo propuesto supone que el producto de la reacción depende linealmente de la temperatura y concentración de la reacción de cierto catalizador. Es probable que se sepa que éste no es el verdadero modelo, sino una aproximación adecuada; de manera que el interés no estribaría en los parámetros individuales, sino en la capacidad de la función en su conjunto para predecir la respuesta verdadera en el rango de las variables consideradas. Por lo tanto, en esta situación, se pondría más énfasis en $\sigma^2_{\hat{y}}$, los intervalos de confianza de la respuesta media, y así sucesivamente, y disminuiría el interés en las inferencias sobre los parámetros individuales.

El experimentador que utiliza análisis de regresión también está interesado en eliminar variables cuando la situación impone que, además de llegar a una ecuación de pronóstico funcional, debe encontrar la "mejor regresión" que implique sólo variables que sean predictores útiles. Se dispone de varios programas de cómputo que llegan en secuencia a la denominada mejor ecuación de regresión, dependiendo de ciertos criterios. En la sección 12.9 profundizaremos en el estudio de esto.

Un criterio que suele utilizarse para ilustrar lo adecuado de un modelo ajustado de regresión es el **coeficiente de determinación múltiple** o $\boldsymbol{R^2}$.

Coeficiente de determinación múltiple o R^2

$$R^2 = \frac{SCR}{STCC} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{SCE}{STCC}.$$

Advierta que esta descripción se parece a la que se hizo de R^2 en el capítulo 11. En este punto la explicación podría ser más clara, toda vez que ahora nos centramos en SCR como la **variabilidad explicada**. La cantidad R^2 tan sólo indica qué proporción de la variación total de la respuesta Y es explicada por el modelo ajustado. Con frecuencia los experimentadores reportan $R^2 \times 100\%$ e interpretan el resultado como el porcentaje de variación explicado con el modelo propuesto. La raíz cuadrada de R^2 se denomina **coeficiente de correlación múltiple** entre Y y el conjunto $x_1, x_2, \ldots, x_k$. En el ejemplo 12.4 el valor de R^2 que indica la proporción de variación explicada por las tres variables independientes x_1, x_2 y x_3 es

$$R^2 = \frac{SCR}{STCC} = \frac{399.45}{438.13} = 0.9117,$$

lo cual significa que 91.17% de la variación del porcentaje de supervivencia queda explicada por el modelo de regresión lineal.

La suma de cuadrados de regresión se puede emplear para obtener algún indicio acerca de si el modelo representa o no una explicación adecuada de la verdadera situación. Podemos probar la hipótesis H_0 de que la **regresión no es significativa** con sólo plantear la razón

$$f = \frac{SCR\ /k}{SCE\ /(n-k-1)} = \frac{SCR\ /k}{s^2}$$

y rechazar H_0 al nivel de significancia α cuando $f > f_\alpha(k, n-k-1)$. Para los datos del ejemplo 12.4 se obtiene

$$f = \frac{399.45/3}{4.298} = 30.98.$$

De la salida de resultados por computadora que aparece en la figura 12.1, el valor P es menor que 0.0001. Esto no debe malinterpretarse. Aunque indica que la regresión explicada por el modelo es significativa, no descarta la posibilidad de que

1. El modelo de regresión lineal en este conjunto de x no sea el único que se puede usar para explicar los datos; de hecho, quizás haya otros modelos con transformaciones sobre las x que generen un valor mayor para el estadístico F.
2. El modelo podría ser más eficaz si se incluyeran otras variables, además de x_1, x_2 y x_3, o quizá si se eliminaran una o más de las variables del modelo, por ejemplo x_3, que tiene un valor $P = 0.5916$.

El lector debería recordar el análisis de la sección 11.5 sobre las desventajas de utilizar R^2 como criterio para comparar modelos en competencia. Es evidente que dichas desventajas son relevantes en la regresión lineal múltiple. De hecho, los riesgos de su empleo en la regresión múltiple son aún mayores debido a que es muy grande la tentación de hacer un sobreajuste. Hay que tener siempre presente que $R^2 \approx 1.0$ siempre puede obtenerse a expensas de los grados de libertad del error cuando se emplea un exceso de términos en el modelo. Sin embargo, $R^2 = 1$, que describe un modelo con ajuste casi perfecto, no siempre genera un modelo que hace buenas predicciones.

El coeficiente de determinación ajustado (R^2_{ajus})

En el capítulo 11 se presentan varias figuras que muestran listados de resultados por computadora, tanto del *SAS* como de *MINITAB*, en las que aparece un estadístico llamado R^2 *ajustado*, o un coeficiente de determinación ajustado. R^2 ajustado es una variación de R^2 que proporciona un **ajuste para los grados de libertad**. El coeficiente de determinación, según se definió en la página 407, no puede disminuir a medida que se agregan términos al modelo. En otras palabras, R^2 no disminuye a medida que se reducen los grados de libertad del error $n - k - 1$, ya que este último resultado se produce por un incremento de k, el número de términos en el modelo. R^2 ajustado se calcula dividiendo la *SCE* y la *STCC* entre sus grados de libertad respectivos de la siguiente manera.

R^2 ajustado

$$R^2_{\text{ajus}} = 1 - \frac{SCE\,/(n-k-1)}{STCC\,/(n-1)}.$$

Para ilustrar el uso de R^2_{ajus} se revisará el ejemplo 12.4.

¿Cómo la eliminación de x_3 afecta a R^2 y R^2_{ajus}?

La prueba t (o la prueba F correspondiente) para x_3 sugiere que un modelo más sencillo que sólo implique x_1 y x_2 bien podría ser una mejoría. En otras palabras, el modelo completo con todos los regresores podría estar sobreajustado. Por supuesto que es interesante investigar R^2 y R^2_{ajus} tanto para el modelo completo (x_1, x_2 y x_3) como para el modelo reducido (x_1, x_2). A partir de la figura 12.1 ya sabemos que $R^2_{\text{compl}} = 0.9117$. La *SCE* para el modelo reducido es 40.01, por lo que $R^2_{\text{redac}} = 1 - \frac{40.01}{438.13} = 0.9087$. De esta forma, con x_3 dentro del modelo se explica más variabilidad. No obstante, como ya se dijo, esto ocurriría aun si el modelo estuviera sobreajustado. Desde luego que R^2_{ajus} está diseñada para proporcionar un estadístico que castigue un modelo sobreajustado, de manera que podríamos esperar que se favorezca al modelo restringido. Entonces, para el modelo completo

$$R^2_{\text{ajus}} = 1 - \frac{38.6764/9}{438.1308/12} = 1 - \frac{4.2974}{36.5109} = 0.8823,$$

mientras que para el modelo reducido (eliminación de x_3)

$$R^2_{\text{ajus}} = 1 - \frac{40.01/10}{438.1308/12} = 1 - \frac{4.001}{36.5109} = 0.8904.$$

Así, R^2_{ajus} realmente favorece el modelo reducido y confirma la evidencia proporcionada por las pruebas t y F, sugiriendo que el modelo reducido es preferible sobre el que contiene los tres regresores. El lector quizás espere que otros estadísticos sugieran el rechazo del modelo sobreajustado. Véase el ejercicio 12.40 de la página 471.

12.7. Caso especial de ortogonalidad

Antes de nuestro desarrollo original del problema general de regresión lineal se planteó la suposición de que las variables independientes se miden sin error y que con frecuencia están bajo el control del experimentador. A menudo ocurren como resultado de un *experimento diseñado con gran detalle*. De hecho, se puede incrementar la eficacia de la ecuación de predicción resultante utilizando un plan de experimentación adecuado.

Suponga que nuevamente consideramos la matriz **X,** tal como se definió en la sección 12.3. Podemos rescribirla como

$$\mathbf{X} = [\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k],$$

donde **1** representa una columna de unos y $\mathbf{x}_j$ es un vector columna que representa los niveles de x_j. Si

$$\mathbf{x}'_p\mathbf{x}_q = \mathbf{0}, \quad \text{para } p \neq q,$$

se dice que las variables x_p y x_q son *ortogonales* entre sí. Hay ciertas ventajas evidentes en tener una situación completamente ortogonal, en la cual $\mathbf{x}'_p\mathbf{x}_q = \mathbf{0}$.

para toda posible p y q, $p \neq q$ y, además,

$$\sum_{i=1}^{n} x_{ji} = 0, \quad j = 1, 2, \ldots, k.$$

La $\mathbf{X'X}$ resultante es una matriz diagonal, y las ecuaciones normales de la sección 12.3 se reducen a

$$nb_0 = \sum_{i=1}^{n} y_i, \quad b_1 \sum_{i=1}^{n} x_{1i}^2 = \sum_{i=1}^{n} x_{1i} y_i, \cdots, b_k \sum_{i=1}^{n} x_{ki}^2 = \sum_{i=1}^{n} x_{ki} y_i.$$

Una ventaja importante es que es fácil hacer la partición de la *SCR* en **componentes de un solo grado de libertad**, cada uno de los cuales corresponde a la cantidad de variación de Y explicada por una variable controlada establecida. En la situación ortogonal se escribe

$$\begin{aligned} SCR &= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n} (b_0 + b_1 x_{1i} + \cdots + b_k x_{ki} - b_0)^2 \\ &= b_1^2 \sum_{i=1}^{n} x_{1i}^2 + b_2^2 \sum_{i=1}^{n} x_{2i}^2 + \cdots + b_k^2 \sum_{i=1}^{n} x_{ki}^2 \\ &= R(\beta_1) + R(\beta_2) + \cdots + R(\beta_k). \end{aligned}$$

La cantidad $R(\beta_i)$ es la cantidad de la suma de cuadrados de regresión asociada con un modelo que implica una sola variable independiente x_i.

Para probar simultáneamente la significancia de un conjunto de m variables en una situación ortogonal, la suma de cuadrados de regresión se convierte en

$$R(\beta_1, \beta_2, \ldots, \beta_m \mid \beta_{m+1}, \beta_{m+2}, \ldots, \beta_k) = R(\beta_1) + R(\beta_2) + \cdots + R(\beta_m),$$

y, por lo tanto,

$$R(\beta_1 \mid \beta_2, \beta_3, \ldots, \beta_k) = R(\beta_1)$$

se simplifica cuando se evalúa una sola variable independiente. Por consiguiente, la contribución de una variable determinada o un conjunto de variables se encuentra, en esencia, *ignorando* las demás variables del modelo. Las evaluaciones independientes del beneficio de las variables individuales se llevan a cabo usando las técnicas de análisis de varianza, tal como se presentan en la tabla 12.4. La variación total en la respuesta está dividida en componentes de un solo grado de libertad más el término del error con $n - k - 1$ grados de libertad. Cada valor f calculado se utiliza para probar una de las hipótesis

$$\left.\begin{aligned} H_0&: \beta_i = 0 \\ H_1&: \beta_i \neq 0 \end{aligned}\right\} \quad i = 1, 2, \ldots, k,$$

comparándolas con el punto crítico $f_\alpha(1, n - k - 1)$ o simplemente interpretando el valor P calculado a partir de la distribución f.

Tabla 12.4: Análisis de varianza para variables ortogonales

Fuente de variación	**Suma de cuadrados**	**Grados de libertad**	**Cuadrados medios**	***f* calculada**
β_1	$R(\beta_1) = b_1^2 \sum_{i=1}^{n} x_{1i}^2$	1	$R(\beta_1)$	$\frac{R(\beta_1)}{s^2}$
β_2	$R(\beta_2) = b_2^2 \sum_{i=1}^{n} x_{2i}^2$	1	$R(\beta_2)$	$\frac{R(\beta_2)}{s^2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
β_k	$R(\beta_k) = b_k^2 \sum_{i=1}^{n} x_{ki}^2$	1	$R(\beta_k)$	$\frac{R(\beta_k)}{s^2}$
Error	SCE	$n-k-1$	$s^2 = \frac{SCE}{n-k-1}$	
Total	$SST = S_{yy}$	$n-1$		

12.8. métodos secuenciales para la Selección del modelo

En ocasiones las pruebas de significancia estudiadas en la sección 12.6 son muy adecuadas para determinar cuáles variables se deben usar en el modelo final de regresión. Dichas pruebas sin duda son eficaces si el experimento se puede planear y las variables son ortogonales entre sí. Incluso si las variables no son ortogonales, las pruebas t individuales se pueden usar en muchos problemas en donde se investigan pocas variables. Sin embargo, existen muchos problemas en los que es necesario utilizar técnicas más elaboradas para seleccionar las variables, en particular si el experimento exhibe una desviación sustancial de la ortogonalidad. Los coeficientes de correlación de la muestra $r_{x_i x_j}$ proporcionan medidas útiles de **multicolinealidad** (dependencia lineal) entre las

variables independientes. Como sólo estamos interesados en la dependencia lineal entre variables independientes, no nos confundiremos si eliminamos las x de la notación y sólo escribimos $r_{x_i x_j} = r_{ij}$, donde

$$r_{ij} = \frac{S_{ij}}{\sqrt{S_{ii} S_{jj}}}.$$

Observe que, en sentido estricto, las r_{ij} no proporcionan estimados verdaderos de los coeficientes de correlación de la población, ya que las x en realidad no son variables aleatorias en el contexto que se estudia aquí. Así, el término *correlación*, aunque estándar, quizá sea inadecuado.

Cuando uno o más de esos coeficientes de correlación muestral se desvía de manera sustancial de cero, suele ser muy difícil encontrar el subconjunto de variables más eficaz para incluirlo en la ecuación de predicción. De hecho, en ciertos problemas la multicolinealidad es tan extrema que no es posible encontrar un predictor adecuado, a menos que se investiguen todos los subconjuntos posibles de variables. En la bibliografía se citan los análisis informativos de Hocking (1976) para la selección de modelos de regresión. En la obra de Myers (1990), también citado, se estudian procedimientos para detectar la multicolinealidad.

El usuario de la regresión lineal múltiple busca lograr uno de tres objetivos:

1. Obtener estimados de coeficientes individuales en un modelo completo.
2. Estudiar variables para determinar cuáles tienen un efecto significativo sobre la respuesta.
3. Calcular la ecuación de predicción más eficaz.

En 1) se sabe de antemano que todas las variables deben incluirse en el modelo. En 2) la predicción es secundaria; mientras que en 3) los coeficientes de regresión individuales no son tan importantes como la calidad de la respuesta estimada $\hat{y}$. Para cada una de las situaciones anteriores la multicolinealidad en el experimento puede tener un efecto profundo sobre el éxito de la regresión.

En esta sección se estudian algunos procedimientos secuenciales estándar para seleccionar variables, los cuales se basan en la idea de que una sola variable o un conjunto de ellas no debería aparecer en la ecuación de estimación, a menos que origine un incremento significativo en la suma de cuadrados de regresión o, en forma equivalente, un incremento significativo de R^2, el coeficiente de determinación múltiple.

12.9. estudio de residuos y violación de suposiciones

Anteriormente en este capítulo se sugirió que los residuales, o errores en el ajuste de regresión, con frecuencia proporcionan información que puede ser muy valiosa para el analista de datos. Los $e_i = y_i = \hat{y}_i$, $i = 1, 2,..., n$, que son el equivalente numérico de los ϵ_i, los errores del modelo, a menudo revelan la posible violación de las suposiciones o la presencia de datos de puntos "sospechosos". Suponga que el vector $\mathbf{x}_i$ denota los valores de las variables regresoras que corresponden al i-ésimo punto de los datos, complementado por un 1 en la posición inicial. Es decir,

$$\mathbf{x}_i' = [1, x_{1i}, x_{2i}, \ldots, x_{ki}].$$

Considere la cantidad

$$h_{ii} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i, \qquad i = 1, 2, ..., n.$$

El lector debería notar que en la sección 12.5 se utilizó h_{ii} para calcular los intervalos de confianza de la respuesta media. Además de σ^2, h_{ii} representa la varianza del valor ajustado $\hat{y}_i$. Los valores h_{ii} son los elementos de la diagonal de la **matriz "SOMBRERO"**

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}',$$

que desempeña un papel importante en cualquier estudio de residuales y en otros aspectos modernos del análisis de regresión (véase Myers, 1990, citado en la bibliografía). El término *matriz SOMBRERO* se deriva del hecho de que **H** genera las "y sombrero", o los valores ajustados cuando se multiplica por el vector **y** de respuestas observadas. Es decir, $\hat{\mathbf{y}} = \mathbf{Xb}$, por lo tanto,

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{Hy},$$

donde $\hat{\mathbf{y}}$ es el vector cuyo i-ésimo elemento es $\hat{y}_i$.

Si se hacen las suposiciones acostumbradas de que los ϵ_i son independientes y están distribuidos normalmente, con media cero y varianza σ^2, las propiedades estadísticas de los residuales se establecen con facilidad. Entonces,

$$E(e_i) = E(y_i - \hat{y}_i) = 0 \quad \text{y} \quad \sigma^2_{e_i} = (1 - h_{ii})\sigma^2,$$

para $i = 1, 2, \ldots, n$. (Para mayores detalles véase Myers, 1990). Es posible demostrar que los valores de la diagonal de la matriz SOMBRERO están acotados de acuerdo con la desigualdad

$$\frac{1}{n} \leq h_{ii} \leq 1.$$

Además, $\sum_{i=1}^{n} h_{ii} = k + 1$, el número de parámetros de la regresión. Como resultado, cualquier punto de los datos cuyo elemento diagonal SOMBRERO sea grande, es decir, esté muy por encima del valor promedio de $(k + 1)/n$, está en una posición dentro del conjunto de datos donde la varianza de $\hat{y}_i$ es relativamente grande y la varianza de un residuo es relativamente pequeña. Como resultado, el analista de datos puede tener una idea de qué tan grande puede ser un residuo antes de que su desviación de cero se pueda atribuir a algo distinto del azar. Muchos de los paquetes comerciales para computadora que permiten calcular la regresión producen el conjunto de **residuales estudentizados**.

Residuo estudentizado

$$r_i = \frac{e_i}{s\sqrt{1 - h_{ii}}}, \quad i = 1, 2, \ldots, n$$

Aquí, cada residuo se **dividió entre una estimación de su desviación estándar** creando un estadístico *tipo t* diseñado para dar al analista una cantidad sin escala que proporcione información sobre el *tamaño* del residual. Además, a menudo los paquetes de cómputo comunes proporcionan valores de otro conjunto de residuales tipo estudentizados denominados **valores *R* de Student**.

Residual *R* de Student

$$t_i = \frac{e_i}{s_{-i}\sqrt{1 - h_{ii}}}, \quad i = 1, 2, \ldots, n,$$

donde s_{-i} es un estimador de la desviación estándar del error calculado con el ***i*-ésimo punto de los datos eliminado**.

Hay tres tipos de transgresiones de las suposiciones fáciles de detectar mediante el uso de los residuales o *gráficas de residuales*. Aunque las gráficas de los residuales brutos, los e_i, ayudan a esto, con frecuencia es más informativo graficar los residuales estudentizados. Las tres transgresiones son las siguientes:

1. Presencia de valores extremos
2. Varianza heterogénea del error
3. Especificación inadecuada del modelo

En el caso 1 elegimos definir un **valor extremo** como un punto de los datos que se desvía de la suposición común de que $E(\epsilon_i) = 0$ para un valor específico de i. Si hay una razón para creer que un punto de un dato específico es un valor extremo que ejerce una gran influencia sobre el modelo ajustado, r_i o t_i, esto podría estar informando algo. Es de esperarse que los valores R de Student sean más sensibles a los valores extremos que los valores r_i.

En realidad, en el caso de que $E(\varepsilon_i) = 0$, t_i es un valor de una variable aleatoria que sigue una distribución t con $n - 1 - (k + 1) = n - k - 2$ grados de libertad. Por consiguiente, es posible utilizar una prueba t de dos colas para proporcionar información con el fin de detectar si el punto i-ésimo es o no un valor extremo.

Aunque el estadístico R de Student t_i produce una prueba t exacta para detectar un valor extremo en una ubicación específica, la distribución t no se aplicaría para probar simultáneamente varios valores extremos en todas las ubicaciones. Como resultado, los residuales estudentizados o valores R de Student se deberían usar estrictamente como herramientas de diagnóstico *sin* un mecanismo de prueba de hipótesis formal. La implicación es que dichos estadísticos resaltan puntos de los datos en los que el error del ajuste es mayor de lo esperado por el azar. Los valores R de Student de gran magnitud sugieren la necesidad de "verificar" los datos con todos los recursos disponibles. La práctica de eliminar observaciones de conjuntos de datos de la regresión no debería llevarse a cabo de forma indiscriminada. (Para más información sobre el uso de los diagnósticos sobre valores extremos véase Myers, 1990, en la bibliografía).

Ilustración de la detección de valores extremos

Estudio de caso 12.1: **Método para capturar saltamontes.** En un experimento biológico, que fue efectuado en el Departamento de Entomología de Virginia Tech, se hicieron n ensayos experimentales con dos métodos diferentes para capturar saltamontes. Los métodos consistieron en la captura por caída de la red y la captura por barrido de la red. El número promedio de saltamontes atrapados con cada método se registró en un conjunto de cuadrantes del campo en una fecha determinada. También se registró una variable regresora adicional, la altura promedio de las plantas en los cuadrantes. Los datos experimentales aparecen en la tabla 12.10.

El objetivo consiste en estimar cuántos saltamontes se capturan empleando sólo el método del barrido de la red, que es menos costoso. Hay cierta preocupación por la validez del cuarto punto de los datos. La captura observada utilizando el método de caída de la red que se reportó parece inusualmente alta, dadas las demás condiciones, de hecho se pensó que la cifra podía ser errónea. Ajuste un modelo del tipo

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

para los 17 puntos de los datos y estudie los residuales para determinar si el punto 4 es un valor extremo.

Tabla 12.10: Conjunto de datos para el estudio de caso 12.1

Observación	Captura por caída de la red, y	Captura por barrido de la red, x_1	Altura de las plantas, x_2 (cm)
1	18.0000	4.15476	52.705
2	8.8750	2.02381	42.069
3	2.0000	0.15909	34.766
4	20.0000	2.32812	27.622
5	2.3750	0.25521	45.879
6	2.7500	0.57292	97.472
7	3.3333	0.70139	102.062
8	1.0000	0.13542	97.790
9	1.3333	0.12121	88.265
10	1.7500	0.10937	58.737
11	4.1250	0.56250	42.386
12	12.8750	2.45312	31.274
13	5.3750	0.45312	31.750
14	28.0000	6.68750	35.401
15	4.7500	0.86979	64.516
16	1.7500	0.14583	25.241
17	0.1333	0.01562	36.354

***Solución*:** Un paquete de cómputo generó el modelo de regresión ajustado

$$\hat{y} = 3.6870 + 4.1050x_1 - 0.0367x_2$$

junto con los estadísticos $R^2 = 0.9244$ y $s^2 = 5.580$. También se obtuvieron los residuales y otra información de diagnóstico que fueron registrados en la tabla 12.11.

Como se esperaba, el residual en la cuarta ubicación parece inusualmente grande, a saber, 7.769. La cuestión fundamental aquí es si este residual es más grande de lo que se esperaría debido al azar. El error estándar del residual para el punto 4 es 2.209. El valor R de Student t_4 que se obtuvo es 9.9315. Al considerarlo como el valor de una variable aleatoria que tiene una distribución t con 13 grados de libertad, se concluiría sin duda que el residuo de la cuarta observación se estima algo mayor que 0, y que la medición del presunto error es apoyada por el estudio de los residuales. Observe que ningún otro valor de los residuales proporciona un valor R de Student que sea motivo de alarma. ∎

Verificación de la normalidad

El lector debe recordar, de acuerdo con lo que se estudió en el capítulo 11, la importancia de verificar la normalidad utilizando una gráfica de probabilidad normal. La misma recomendación es válida para el caso de la regresión lineal múltiple. Las gráficas de probabilidad normal se pueden generar utilizando software estándar para regresión. Sin embargo, como ya se indicó, éstas pueden ser más eficaces si se usan residuales estudentizados o valores R de Student en vez de residuales comunes.

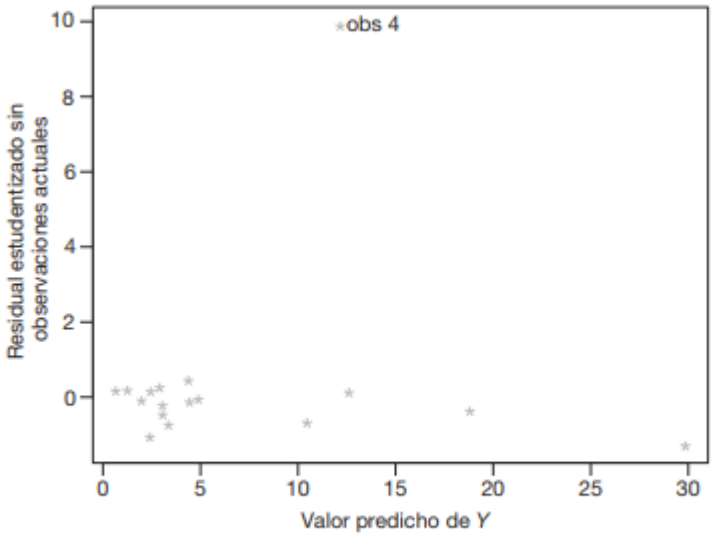


Figura 12.5: Valores R de Student graficados en comparación con los valores predichos para los datos de los saltamontes del estudio de caso 12.1.

12.10. validación Cruzada C y otros criterios para la Selección del modelo

Para muchos problemas de regresión el experimentador debe elegir entre varios modelos alternativos o formas de modelo que se desarrollan a partir del mismo conjunto de datos. Con mucha frecuencia se requiere el modelo que predice o estima mejor la respuesta media. El experimentador debe tomar en cuenta los tamaños relativos de los valores de s^2 para los posibles modelos y, sin duda, la naturaleza general de los intervalos de confianza sobre la respuesta media. También se debe considerar lo bien que el modelo predice los valores de la respuesta que **no se hayan utilizado para construir los posibles modelos**. Los modelos deben estar sujetos a **validación cruzada**. Entonces, lo que se requiere son los errores de la validación cruzada en lugar de los errores del ajuste. Estos errores en la predicción son los **residuales PRESS**.

$$\delta_i = y_i - \hat{y}_{i,-i}, \quad i = 1, 2, \ldots, n,$$

donde $\hat{y}_{i,-i}$ es la predicción del i-ésimo punto de los datos por medio de un modelo que no utiliza el i-ésimo punto en el cálculo de los coeficientes. Estos residuales PRESS se calculan mediante la fórmula

$$\delta_i = \frac{e_i}{1 - h_{ii}}, \quad i = 1, 2, \ldots, n.$$

(La derivación se encuentra en Myers, 1990).

Uso del estadístico PRESS

La motivación para utilizar PRESS y la utilidad de los residuales PRESS es muy fácil de entender. El propósito de extraer o *separar* puntos de datos, uno a la vez, consiste en

permitir el empleo de metodologías separadas para ajustar y evaluar un modelo específico. Para evaluar un modelo la "$-i$" indica que el residual PRESS proporciona un error de predicción donde la observación que se predice es *independiente del ajuste del modelo*.

Los criterios que utilizan los residuales PRESS son dados por

$$\sum_{i=1}^{n} |\delta_i| \quad \text{y} \quad \text{PRESS} = \sum_{i=1}^{n} \delta_i^2.$$

(El término **PRESS** es un acrónimo que se forma con las iniciales de los términos de la frase en inglés *prediction sum of squares,* que se traduce como **suma de cuadrados de predicción**). Se sugiere que se utilicen ambos criterios. Es posible que PRESS sea dominado por uno o algunos residuales PRESS grandes. Es evidente que el criterio sobre $\sum_{i=1}^{n}|\delta_i|$ es menos sensible a un número pequeño de valores grandes.

Además del estadístico PRESS en sí, el analista puede simplemente calcular un estadístico similar a R^2 que refleje el desempeño de la predicción. Con frecuencia a este estadístico se le denomina R^2_{pred} y se calcula como sigue:

R^2 de predicción

Dado un modelo ajustado con valor específico para PRESS, R^2_{pred} es dado por

$$R^2_{\text{pred}} = 1 - \frac{\text{PRESS}}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

Observe que R^2_{pred} es tan sólo el estadístico común R^2 donde la *SCE* fue reemplazada por el estadístico PRESS.

En el siguiente estudio de caso se proporciona un ejemplo en el que se ajustan muchos posibles modelos a un conjunto de datos y se elige el mejor de ellos. No se emplean los procedimientos secuenciales descritos en la sección 12.9. En vez de eso se ilustra el papel que desempeñan los residuales PRESS y otros valores estadísticos cuando se trata de seleccionar la mejor ecuación de regresión.

Estudio de caso 12.2:

Patada de fútbol. La fuerza de las piernas es un requisito necesario para que un pateador tenga éxito en el fútbol americano. Una medida de la calidad de una buena patada es el "tiempo de vuelo" del balón, es decir, el tiempo que el balón se mantiene en el aire antes de ser atrapado por el regresador de patadas. Para determinar cuáles factores de la fuerza de las piernas influyen en el tiempo de vuelo y desarrollar un modelo empírico para predecir esta respuesta, el Departamento de Salud, Educación Física y Recreación de Virginia Tech llevó a cabo un estudio sobre *La relación entre variables seleccionadas de desempeño físico y la capacidad de despejes en el fútbol*. Se eligieron 13 pateadores para el experimento y cada uno pateó 10 veces el balón. En la tabla 12.12 aparece el registro del tiempo de vuelo promedio, junto con las medidas de fuerza usadas en el análisis.

Cada variable regresora se define como sigue:

1. **FPD**, fuerza de la pierna derecha (libras)
2. **FPI**, fuerza de la pierna izquierda (libras)
3. **FTD**, flexibilidad muscular del tendón derecho (grados)
4. **FTI**, flexibilidad muscular del tendón izquierdo (grados)

5. Potencia, fuerza general de las piernas (pie-libras)

Determine el modelo más adecuado para predecir el tiempo de vuelo.

Tabla 12.12: Datos para el estudio de caso 12.2.

Pateador	Tiempo de vuelo, y (seg)	FPD, x_1	FPI, x_2	FTD, x_3	FTI, x_4	Potencia, x_5
1	4.75	170	170	106	106	240.57
2	4.07	140	130	92	93	195.49
3	4.04	180	170	93	78	152.99
4	4.18	160	160	103	93	197.09
5	4.35	170	150	104	93	266.56
6	4.16	150	150	101	87	260.56
7	4.43	170	180	108	106	219.25
8	3.20	110	110	86	92	132.68
9	3.02	120	110	90	86	130.24
10	3.64	130	120	85	80	205.88
11	3.68	120	140	89	83	153.92
12	3.60	140	130	92	94	154.64
13	3.85	160	150	95	95	240.57

***Solución*:** Al buscar el mejor modelo posible para predecir el tiempo de vuelo se obtuvo la información de la tabla 12.13 utilizando un paquete de cómputo para regresión. Los modelos están clasificados en orden ascendente con respecto a los valores del estadístico PRESS. Esta presentación brinda información suficiente acerca de todos los modelos posibles con el fin de permitir que el usuario elimine algunos de ellos. Al parecer, el mejor modelo para predecir el tiempo de vuelo para los pateadores es el que contiene a x_2 y x_5 (FPI y potencia), denotadas por x_2x_5. Asimismo, observe que todos los modelos con valores bajos de PRESS, de s^2, de $\sum_{i=1}^{n}|\delta_i|$, y con valores altos de R^2, contienen esas dos variables. Para obtener información de los residuales de la regresión ajustada

$$\hat{y}_i = b_0 + b_2x_{2i} + b_5x_{5i},$$

se generaron los residuales y los residuales PRESS. El modelo de predicción real (véase el ejercicio 12.47 de la página 494) es dado por

$$\hat{y} = 1.10765 + 0.01370x_2 + 0.00429x_5.$$

En la tabla 12.14 se listan los residuales, los valores de la diagonal testada y los valores PRESS.

Observe el ajuste relativamente bueno de los modelos de regresión con dos variables para los datos. Los residuales PRESS reflejan la capacidad de la ecuación de regresión para predecir el tiempo de vuelo si se hicieran predicciones independientes. Por ejemplo, para el pateador número 4 el tiempo de vuelo de 4.180 tendría un error de predicción de 0.039 si se construyera el modelo usando a los 12 pateadores restantes. Para este modelo el error promedio de la predicción, o error de validación cruzada, es

$$\frac{1}{13}\sum_{i=1}^{n}|\delta_i| = 0.1489 \text{ segundos},$$

Tabla 12.13: Comparación de diferentes modelos de regresión

Modelo	s^2	$\sum \lvert\delta_i\rvert$	**PRESS**	R^2
x_2x_5	0.036907	1.93583	0.54683	0.871300
$x_1x_2x_5$	0.041001	2.06489	0.58998	0.871321
$x_2x_4x_5$	0.037708	2.18797	0.59915	0.881658
$x_2x_3x_5$	0.039636	2.09553	0.66182	0.875606
$x_1x_2x_4x_5$	0.042265	2.42194	0.67840	0.882093
$x_1x_2x_3x_5$	0.044578	2.26283	0.70958	0.875642
$x_2x_3x_4x_5$	0.042421	2.55789	0.86236	0.881658
$x_1x_3x_5$	0.053664	2.65276	0.87325	0.831580
$x_1x_4x_5$	0.056279	2.75390	0.89551	0.823375
x_1x_5	0.059621	2.99434	0.97483	0.792094
x_2x_3	0.056153	2.95310	0.98815	0.804187
x_1x_3	0.059400	3.01436	0.99697	0.792864
$x_1x_2x_3x_4x_5$	0.048302	2.87302	1.00920	0.882096
x_2	0.066894	3.22319	1.04564	0.743404
x_3x_5	0.065678	3.09474	1.05708	0.770971
x_1x_2	0.068402	3.09047	1.09726	0.761474
x_3	0.074518	3.06754	1.13555	0.714161
$x_1x_3x_4$	0.065414	3.36304	1.15043	0.794705
$x_2x_3x_4$	0.062082	3.32392	1.17491	0.805163
x_2x_4	0.063744	3.59101	1.18531	0.777716
$x_1x_2x_3$	0.059670	3.41287	1.26558	0.812730
x_3x_4	0.080605	3.28004	1.28314	0.718921
x_1x_4	0.069965	3.64415	1.30194	0.756023
x_1	0.080208	3.31562	1.30275	0.692334
$x_1x_3x_4x_5$	0.059169	3.37362	1.36867	0.834936
$x_1x_2x_4$	0.064143	3.89402	1.39834	0.798692
$x_3x_4x_5$	0.072505	3.49695	1.42036	0.772450
$x_1x_2x_3x_4$	0.066088	3.95854	1.52344	0.815633
x_5	0.111779	4.17839	1.72511	0.571234
x_4x_5	0.105648	4.12729	1.87734	0.631593
x_4	0.186708	4.88870	2.82207	0.283819

que es pequeño comparado con el tiempo de vuelo promedio para los 13 pateadores. ∎

En la sección 12.9 indicamos que a menudo es aconsejable utilizar todos los subconjuntos posibles de regresión cuando se busca el mejor modelo. La mayoría de los programas comerciales de cómputo para estadística contienen una rutina de *todas las regresiones posibles*. Tales algoritmos calculan diversos criterios para todos los subconjuntos de términos del modelo. Es evidente que criterios como R^2, s^2 y PRESS son razonables para elegir entre subconjuntos de candidatos. Otro estadístico muy popular y útil, en particular para las ciencias físicas e ingeniería, es el estadístico C_p, que se describe a continuación.

Tabla 12.14: Residuales PRESS

Pateador	y_i	$\hat{y}_i$	$e_i = y_i - \hat{y}_i$	h_{ii}	δ_i
1	4.750	4.470	0.280	0.198	0.349
2	4.070	3.728	0.342	0.118	0.388
3	4.040	4.094	−0.054	0.444	−0.097
4	4.180	4.146	0.034	0.132	0.039
5	4.350	4.307	0.043	0.286	0.060
6	4.160	4.281	−0.121	0.250	−0.161
7	4.430	4.515	−0.085	0.298	−0.121
8	3.200	3.184	0.016	0.294	0.023
9	3.020	3.174	−0.154	0.301	−0.220
10	3.640	3.636	0.004	0.231	0.005
11	3.680	3.687	−0.007	0.152	−0.008
12	3.600	3.553	0.047	0.142	0.055
13	3.850	4.196	−0.346	0.154	−0.409

El estadístico C_p

Muy a menudo la selección del modelo más adecuado implica tomar en cuenta muchas cosas. Evidentemente el número de términos del modelo es importante; el tema de la parsimonia no debe ignorarse. Por otro lado, el analista no debe sentirse satisfecho con un modelo demasiado simple hasta el punto de una simplificación excesiva. En este sentido, un estadístico único que implica un compromiso aceptable es C_p. (Véase Mallows, 1973, en la bibliografía).

El estadístico C_p apela de forma muy adecuada al sentido común y se desarrolla tomando en cuenta el equilibrio apropiado entre el sesgo excesivo en que se incurre cuando se subajusta, es decir, cuando se eligen muy pocos términos para el modelo; y la varianza excesiva de la predicción que se genera cuando se sobreajusta, o sea cuando hay redundancias en el modelo. El estadístico C_p es una función simple del número total de parámetros en el posible modelo y la media cuadrada del error s^2.

Aquí no presentaremos el desarrollo completo del estadístico C_p. (Para mayores detalles se recomienda consultar a Myers, 1990, listado en la bibliografía). El C_p para un subconjunto particular de modelos es *un estimado* de lo siguiente:

$$\Gamma_{(p)} = \frac{1}{\sigma^2}\sum_{i=1}^{n} \mathrm{Var}(\hat{y}_i) + \frac{1}{\sigma^2}\sum_{i=1}^{n} (\mathrm{Sesgo}\,\hat{y}_i)^2.$$

Se descubre que, bajo las suposiciones estándar de los mínimos cuadrados que se indicaron con anterioridad en este capítulo, y asumiendo que el modelo "verdadero" es aquel que contiene todas las posibles variables,

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} \mathrm{Var}(\hat{y}_i) = p \qquad \text{(número de parámetros en el posible modelo)}$$

(véase el ejercicio de repaso 12.63) y un estimado no sesgado de

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(\text{Sesgo}\,\hat{y}_i)^2 \text{ es dado por } \frac{1}{\sigma^2}\sum_{i=1}^{n}(\widehat{\text{Sesgo}}\,\hat{y}_i)^2 = \frac{(s^2-\sigma^2)(n-p)}{\sigma^2}.$$

En las ecuaciones anteriores s^2 es el cuadrado medio del error para el posible modelo y σ^2 es la varianza del error de la población. Así, si asumimos que se dispone de algún estimado $\hat{\sigma}^2$ para σ^2, entonces C_p es dado por la siguiente ecuación:

Estadístico C_p

$$C_p = p + \frac{(s^2-\hat{\sigma}^2)(n-p)}{\hat{\sigma}^2},$$

donde p es el número de parámetros en el modelo, s^2 es el cuadrado medio del error para el modelo candidato y $\hat{\sigma}^2$ es un estimador de σ^2.

Es evidente que el científico debería adoptar modelos con valores pequeños de C_p. El lector observará que, a diferencia del estadístico PRESS, C_p carece de una escala. Además, se puede obtener cierta información acerca de qué tan adecuado es un posible modelo observando su valor de C_p. Por ejemplo, $C_p > p$ indica que un modelo está sesgado debido a que está subajustado, mientras que $C_p \approx p$ indica un modelo razonable.

Con frecuencia hay confusión respecto a la procedencia de $\hat{\sigma}^2$ en la fórmula para C_p. Es evidente que el científico o ingeniero no tienen acceso a la cantidad σ^2 de la población. En aplicaciones donde se dispone de corridas repetidas, digamos en situaciones de diseño experimental, se dispone de un estimado de σ^2 independiente del modelo (véase los capítulos 11 y 15). Sin embargo, la mayoría de paquetes de cómputo utilizan $\hat{\sigma}^2$ como el *cuadrado medio del error del modelo más completo*. Evidentemente, si éste no es un buen estimado, la parte de sesgo del estadístico C_p puede ser negativa. Por consiguiente, C_p puede ser menor que p.

13. EXPERIMENTOS DE UN FACTOR GENERAL

13.2.técnica de análisis de varianza

En el material sobre estimación y prueba de hipótesis que se cubrió en los capítulos 9 y 10 en cada caso nos limitamos a considerar sólo dos parámetros de la población. Ése fue el caso, por ejemplo, en la prueba de la igualdad de dos medias de la población, en la cual se usaron muestras independientes de poblaciones normales con varianza común pero desconocida, y en donde se necesitaba obtener un estimado agrupado de σ^2.

El material que se refiere a las inferencias de dos muestras representa un caso especial de lo que se denomina *problema de un solo factor*. Por ejemplo, en el ejercicio 10.35 de la página 357 se midió el tiempo de supervivencia para dos muestras de ratones, en donde una muestra recibió un tratamiento de suero contra la leucemia y la otra no lo recibió. En este caso decimos que hay *un factor*, es decir, un *tratamiento*, y el factor se halla en *dos niveles*. Si en el proceso de muestreo se utilizaran varios tratamientos en competencia, se necesitarían más muestras de ratones. En ese caso el problema implicaría un factor con más de dos niveles, por lo tanto, con más de dos muestras.

En el problema de $k > 2$ muestras se supone que hay k muestras provenientes de k poblaciones. Un procedimiento muy común que se utiliza cuando se prueban medias de la población se denomina **análisis de varianza**, o **ANOVA**.

Si el lector ha estudiado el material acerca de la teoría de la regresión, el análisis de varianza no será, por supuesto, una técnica nueva para él. Utilizamos el método del análisis de varianza para partir la suma total de cuadrados en dos partes, una parte debida a la regresión y otra debida al error.

Suponga que en un experimento industrial a un ingeniero le interesa la forma en que la absorción media de humedad del concreto varía para 5 agregados de concreto diferentes. Las muestras se exponen a la humedad durante 48 horas y se decide que para cada agregado deben probarse 6 muestras, lo que hace que se requiera probar un total de 30 muestras. En la tabla 13.1 se presentan los datos registrados.

El modelo que se considera para esta situación es el siguiente. Se tomaron 6 observaciones de cada una de las 5 poblaciones, con medias $\mu_1, \mu_2, \ldots, \mu_5$, respectivamente. Deseamos probar

$$H_0: \ \mu_1 = \mu_2 = \cdots = \mu_5,$$
$$H_1: \ \text{Al menos dos de las medias no son iguales.}$$

Tabla 13.1: Absorción de humedad en agregados para concreto

Agregado:	**1**	**2**	**3**	**4**	**5**	
	551	595	639	417	563	
	457	580	615	449	631	
	450	508	511	517	522	
	731	583	573	438	613	
	499	633	648	415	656	
	632	517	677	555	679	
Total	3320	3416	3663	2791	3664	16,854
Media	553.33	569.33	610.50	465.17	610.67	561.80

Además, estamos interesados en realizar comparaciones individuales entre estas 5 medias de la población.

Dos fuentes de variabilidad en los datos

En el procedimiento del análisis de varianza se supone que cualquier variación que exista entre los promedios de los agregados se atribuye a 1) la variación en la absorción entre observaciones *dentro* de los tipos de agregados, y 2) la variación *entre* los tipos de agregados, es decir, a las diferencias en la composición química de los agregados. Por supuesto, la **variación dentro de los agregados** se debe a varias causas. Quizá las condiciones de temperatura y humedad no se mantuvieron constantes durante el experimento. Es posible que haya habido cierta cantidad de heterogeneidad en los lotes de materias primas que se usaron. En todo caso debe considerarse la variación dentro de la muestra como una **variación aleatoria o al azar**. Parte del objetivo del análisis de varianza consiste en determinar si las diferencias entre las 5 medias muestrales son lo que se esperaría debido sólo a la variación aleatoria o si, más bien, se trata de una variación más allá de los simples efectos del azar, como las diferencias en la composición química de los agregados.

En esta etapa surgen muchas preguntas acerca del problema anterior. Por ejemplo, ¿cuántas muestras deben probarse para cada agregado? Ésta es una pregunta que desafía continuamente al analista. Además, ¿qué pasaría si la variación dentro de la muestra fuera tan grande que al procedimiento estadístico le resultara difícil detectar las diferencias sistemáticas? ¿Es posible controlar de manera sistemática fuentes externas de variación y así eliminarlas de la parte que llamamos variación aleatoria? En las secciones siguientes intentaremos responder éstas y otras preguntas.

13.3. la estrategia del diseño experimental

En los capítulos 9 y 10 se estudiaron los conceptos de la estimación y la prueba de hipótesis para el caso de dos muestras, bajo la importante perspectiva de la manera en que se realiza el experimento. Esto forma parte de la categoría amplia de los diseños experimentales. Por ejemplo, para la **prueba *t* agrupada** que se estudió en el capítulo 10, se supone que los niveles de los factores (los tratamientos en el ejemplo de los ratones) se asignan al azar a las unidades experimentales (los ratones). En los capítulos 9 y 10 analizamos el

concepto de unidades experimentales y lo ilustramos por medio de varios ejemplos. En pocas palabras, las unidades experimentales son las unidades (ratones, pacientes, especímenes de concreto, tiempo) que **proporcionan la heterogeneidad que conduce al error experimental** en una investigación científica. La asignación aleatoria elimina el sesgo que podría originarse con una asignación sistemática. El objetivo consiste en distribuir en forma uniforme entre los niveles de los factores los riesgos que introduce la heterogeneidad de las unidades experimentales. Una asignación al azar simula mejor las condiciones que se asumen en el modelo. En la sección 13.7 analizamos la **formación de bloques** en los experimentos. En los capítulos 9 y 10 se presentó el concepto, cuando se efectuaron comparaciones entre las medias usando el **emparejamiento**, es decir, la división de las unidades experimentales en pares homogéneos denominados **bloques**. Entonces, los niveles de los factores o tratamientos se asignan al azar dentro de los bloques. El propósito de la formación de bloques es reducir el error experimental efectivo. En este capítulo se extiende de manera natural el emparejamiento a bloques de tamaño mayor, con el análisis de varianza como la herramienta analítica principal.

13.4. análisis de varianza unilateral diseño completamente aleatorio

De k poblaciones se seleccionan muestras aleatorias de tamaño n. Las k poblaciones diferentes se clasifican con base en un criterio único, como tratamientos o grupos distintos. En la actualidad el término **tratamiento** se utiliza por lo general para designar las diversas clasificaciones, ya sean diferentes agregados, analistas, fertilizadores o regiones del país.

Suposiciones e hipótesis del ANOVA de un solo factor

Se supone que las k poblaciones son independientes y que están distribuidas en forma normal con medias $\mu_1, \mu_2, \ldots, \mu_k$, y varianza común σ^2. Como se indicó en la sección 13.2, estas suposiciones son más aceptables mediante la aleatoriedad. Se desean obtener métodos adecuados para probar las hipótesis

$$H_0: \ \mu_1 = \mu_2 = \cdots = \mu_k,$$

$$H_1: \ \text{Al menos dos de las medias no son iguales.}$$

Sea que y_{ij} denote la j-ésima observación del i-ésimo tratamiento, y el acomodo de los datos es el que se observa en la tabla 13.2. Aquí, Y_i es el total de todas las observaciones de la muestra, del i-ésimo tratamiento, $\bar{y}_i$, es la media de todas las observaciones en la muestra del i-ésimo tratamiento, $Y_{..}$ es el total de todas las nk observaciones, y $\bar{y}_{...}$ es la media de todas las nk observaciones.

Modelo de ANOVA para un solo factor

Cada observación puede escribirse en la forma

$$Y_{ij} = \mu_i + \epsilon_{ij},$$

donde ϵ_{ij} mide la desviación que tiene la observación j-ésima de la i-ésima muestra, con respecto de la media del tratamiento correspondiente. El término ϵ_{ij} representa el error aleatorio y desempeña el mismo papel que los términos del error en los modelos de

Tabla 13.2: k muestras aleatorias

Tratamiento:	**1**	**2**	$\cdots$	i	$\cdots$	k	
	y_{11}	y_{21}	$\cdots$	y_{i1}	$\cdots$	y_{k1}	
	y_{12}	y_{22}	$\cdots$	y_{i2}	$\cdots$	y_{k2}	
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	
	y_{1n}	y_{2n}	$\cdots$	y_{in}	$\cdots$	y_{kn}	
Total	$Y_{1.}$	$Y_{2.}$	$\cdots$	$Y_{i.}$	$\cdots$	$Y_{k.}$	$Y_{..}$
Media	$\bar{y}_{1.}$	$\bar{y}_{2.}$	$\cdots$	$\bar{y}_{i.}$	$\cdots$	$\bar{y}_{k.}$	$\bar{y}_{..}$

regresión. Una forma alternativa y preferible de esta ecuación se obtiene sustituyendo $\mu_i = \mu + \alpha_i$, sujeta a la restricción $\sum_{i=1}^{k} \alpha_i = 0$. Por lo tanto, se escribe

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

donde μ tan sólo es la **media general** de todas las μ_i, es decir,

$$\mu = \frac{1}{k}\sum_{i=1}^{k} \mu_i,$$

y α_i se denomina el **efecto** del i-ésimo tratamiento.

La hipótesis nula de que k medias de la población son iguales, en comparación con la alternativa de que al menos dos de las medias son distintas, ahora se puede reemplazar por las hipótesis equivalentes.

$$H_0: \ \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0,$$
$$H_1: \text{Al menos una de las } \alpha_i \text{ no es igual a cero.}$$

Resolución de la variabilidad total en componentes

Nuestra prueba se basará en una comparación de dos estimados independientes de la varianza poblacional común σ^2. Dichos estimadores se obtendrán haciendo la partición de la variabilidad total de nuestros datos, denotados mediante la sumatoria doble

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2,$$

en dos componentes.

Teorema 13.1: Identidad de la suma de cuadrados

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2 = n\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2$$

En lo que sigue, será conveniente identificar los términos de la identidad de la suma de cuadrados con la siguiente notación:

Tres medidas importantes de variabilidad

$$STC = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2 = \text{suma total de cuadrados,}$$

$$SCT = n\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..})^2 = \text{suma de los cuadrados del tratamiento,}$$

$$SCE = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2 = \text{suma de los cuadrados del error.}$$

Entonces, la identidad de la suma de los cuadrados se puede representar simbólicamente con la ecuación

$$STC = SCT + SCE$$

La identidad anterior expresa cómo las variaciones entre los tratamientos y dentro de los tratamientos contribuyen a la suma total de cuadrados. Sin embargo, se puede obtener mucha información si se investiga el **valor esperado tanto de *SCT* como de *SCE***. Eventualmente calcularemos estimados de la varianza que determinan la razón que

Teorema 13.2:

$$E(SCT) = (k-1)\sigma^2 + n\sum_{i=1}^{k}\alpha_i^2$$

La prueba del teorema se deja como ejercicio para el lector (véase el ejercicio 13.53 de la página 556).

Si H_0 es verdadera, un estimado de σ^2 basado en $k-1$ grados de libertad es dado por la expresión:

Media cuadrática del tratamiento

$$s_1^2 = \frac{SCT}{k-1}$$

Si H_0 es verdadera y por ello cada α_i en el teorema 13.2 es igual a cero, se observa que

$$E\left(\frac{SCT}{k-1}\right) = \sigma^2,$$

y s_1^2 es un estimado no sesgado de σ^2. Sin embargo, si H_1 es verdadera, se tiene que

$$E\left(\frac{SCT}{k-1}\right) = \sigma^2 + \frac{n}{k-1}\sum_{i=1}^{k}\alpha_i^2,$$

y s_1^2 estima a σ^2 más un término adicional, que mide la variación debida a los efectos sistemáticos.

Otro estimado independiente de σ^2, basado en $k(n-1)$ grados de libertad, es la fórmula familiar:

Cuadrado medio del error

$$s^2 = \frac{SCE}{k(n-1)}$$

13.5. pruebas de la igualdad de varias varianzas

Resulta aleccionador puntualizar la importancia de los valores esperados de los cuadrados medios a los que recién nos referimos. En la sección siguiente se estudia el empleo de una **razón F** con el cuadrado medio del tratamiento en el numerador. Se observa que cuando H_1 es verdadera, la presencia de la condición $E(s_1^2) > E(s^2)$ sugiere que la razón F se utiliza en el contexto de una **prueba unilateral de cola superior**. Es decir, cuando H_1 es verdadera se esperaría que el numerador s_1^2 fuera mayor que el denominador.

Uso de la prueba F en el ANOVA

El estimado s^2 es no sesgado, independientemente de la veracidad o falsedad de la hipótesis nula (véase el ejercicio de repaso 13.52 de la página 556). Es importante señalar que la identidad de la suma de cuadrados ha hecho la partición no sólo de la variabilidad total de los datos, sino también del número total de grados de libertad. Es decir,

$$nk - 1 = k - 1 + k(n - 1).$$

Razón F para probar la igualdad de las medias

Cuando H_0 es verdadera, la razón $f = s_1^2/s^2$ es un valor de la variable aleatoria F, que tiene la distribución F con $k - 1$ y $k(n - 1)$ grados de libertad (véase el teorema 8.8). Como s_1^2 sobrestima a σ^2 cuando H_0 es falsa, se tiene una prueba de una cola con la región crítica localizada por completo en la cola derecha de la distribución.

A un nivel de significancia de α se rechaza la hipótesis nula H_0 cuando

$$f > f_\alpha[k - 1, k(n - 1)].$$

Otro método, el del valor P, sugiere que la evidencia a favor o en contra de H_0 es

$$P = P\{f[k - 1,\ k(n - 1)] > f\}.$$

Los cálculos para un problema de análisis de varianza por lo general se resumen en forma tabular, como se observa en la tabla 13.3.

Tabla 13.3: Análisis de varianza para el ANOVA de un solo factor

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	f calculada
Tratamientos	SCT	$k - 1$	$s_1^2 = \dfrac{SCT}{k - 1}$	$\dfrac{s_1^2}{s^2}$
Error	SCE	$k(n - 1)$	$s^2 = \dfrac{SCE}{k(n - 1)}$	
Total	STC	$kn - 1$		

13.6. Comparaciones de un solo grado de libertad

El análisis de varianza en la clasificación de un solo factor, o experimento de un solo factor, como se le denomina con frecuencia, tan sólo indica si puede rechazarse o no la hipótesis de medias de tratamientos iguales. Por lo general, el experimentador preferiría efectuar un análisis más profundo. Como ilustración, en el ejemplo 13.1, mediante el rechazo de la hipótesis nula, concluimos que las medias no son iguales, pero aún no sabemos en dónde residen las diferencias entre los agregados. Es probable que el ingeniero intuya de antemano que los agregados 1 y 2 deberían poseer propiedades similares de absorción, al igual que los agregados 3 y 5. Sin embargo, sería interesante estudiar las diferencias entre los dos grupos. Así, parece apropiado probar las hipótesis

$$H_0\colon\ \mu_1 + \mu_2 - \mu_3 - \mu_5 = 0,$$
$$H_1\colon\ \mu_1 + \mu_2 - \mu_3 - \mu_5 \neq 0.$$

Se observa que la hipótesis es una función lineal de las medias de la población, en las cuales los coeficientes suman cero.

Definición 13.1: Cualquier función lineal de la forma

$$\omega = \sum_{i=1}^{k} c_i \mu_i,$$

donde $\sum_{i=1}^{k} c_i = 0$ se llama **comparación** o **contraste** en las medias de los tratamientos.

Con frecuencia el experimentador puede hacer comparaciones múltiples probando la significancia de los contrastes de las medias de los tratamientos, es decir, probando una hipótesis del siguiente tipo:

Estadístico de prueba para probar un contraste

Nuestra hipótesis se prueba a un nivel de significancia α calculando

$$f = \frac{\left(\sum_{i=1}^{k} c_i \bar{y}_{i.}\right)^2}{s^2 \sum_{i=1}^{k} (c_i^2/n_i)} = \frac{\left[\sum_{i=1}^{k} (c_i Y_{i.}/n_i)\right]^2}{s^2 \sum_{i=1}^{k} (c_i^2/n_i)} = \frac{SCw}{s^2}.$$

Aquí f es un valor de la variable aleatoria F que tiene distribución F con 1 y $N-k$ grados de libertad.

Cuando los tamaños de las muestras son iguales a n,

$$SCw = \frac{\left(\sum_{i=1}^{k} c_i Y_{i.}\right)^2}{n \sum_{i=1}^{k} c_i^2}.$$

La cantidad SCw, que se denomina **suma de cuadrados de los contrastes**, indica la parte de la SCT que se explica por el contraste en cuestión.

13.7. comparaciones múltiples

Esta suma de cuadrados se empleará para probar la hipótesis de que

$$\sum_{i=1}^{k} c_i \mu_i = 0.$$

Con frecuencia es de interés probar contrastes múltiples, en particular contrastes que son linealmente independientes u ortogonales. Como resultado, se vuelve necesaria la siguiente definición:

Definición 13.2: Se dice que los dos contrastes

$$\omega_1 = \sum_{i=1}^{k} b_i \mu_i \qquad \text{y} \qquad \omega_2 = \sum_{i=1}^{k} c_i \mu_i$$

son **ortogonales**, si $\sum_{i=1}^{k} b_i c_i / n_i = 0$, o bien, cuando las n_i son iguales a n, si

$$\sum_{i=1}^{k} b_i c_i = 0.$$

Si ω_1 y ω_2 son ortogonales, entonces las cantidades SCw_1 y SCw_2 son componentes de SCT, cada una con un solo grado de libertad. La suma de cuadrados de los tratamientos con $k-1$ grados de libertad se puede dividir en, a lo sumo, $k-1$ sumas de cuadrados de contrastes independientes con un solo grado de libertad que satisfacen la identidad

$$SCT = SCw_1 + SCw_2 + \cdots + SCw_{k-1},$$

si los contrastes son ortogonales entre sí.

13.8. comparación de tratamientos con un control

El análisis de varianza es un procedimiento poderoso para probar la homogeneidad de un conjunto de medias. No obstante, si se rechazara la hipótesis nula y se aceptara la alternativa que se planteó (que no todas las medias son iguales), aún no se sabría cuáles de las medias de la población son iguales y cuáles son diferentes.

```
                              The GLM Procedure
Dependent Variable: moisture
                                  Sum of
 Source                DF        Squares     Mean Square    F Value    Pr > F
 Model                  4     85356.4667      21339.1167       4.30    0.0088
 Error                 25    124020.3333       4960.8133
 Corrected Total       29    209376.8000

          R-Square      Coeff Var       Root MSE     moisture Mean
          0.407669       12.53703       70.43304          561.8000

 Source               DF      Type I SS      Mean Square      F Value      Pr > F
 aggregate             4    85356.46667      21339.11667         4.30      0.0088

 Source               DF    Type III SS      Mean Square      F Value      Pr > F
 aggregate             4    85356.46667      21339.11667         4.30      0.0088

 Contrast             DF   Contrast SS      Mean Square      F Value      Pr > F
 (1,2,3,5) vs. 4       1    70035.00833      70035.00833        14.12      0.0009
 (1,2) vs. (3,5)       1    14553.37500      14553.37500         2.93      0.0991
 1 vs. 2             1        768.00000        768.00000         0.15      0.6973
 3 vs. 5             1          0.08333          0.08333         0.00      0.9968
```

Figura 13.4: Un conjunto de contrastes ortogonales.

A menudo es de interés efectuar varias **comparaciones por pares** (quizá todas las que sean posibles) entre los tratamientos. En realidad, una comparación por pares se puede ver como un contraste simple, es decir, una prueba de

$$H_0\colon\ \mu_i - \mu_j = 0,$$
$$H_1\colon\ \mu_i - \mu_j \neq 0,$$

para toda $i \neq j$. Hacer todas las comparaciones posibles por pares entre las medias puede ser muy benéfico cuando no se conocen *a priori* contrastes complejos particulares. Por ejemplo, suponga que se desea probar las hipótesis siguientes, con los datos de los agregados de la tabla 13.1:

$$H_0\colon\ \mu_1 - \mu_5 = 0,$$
$$H_1\colon\ \mu_1 - \mu_5 \neq 0.$$

La prueba se desarrolla usando una F, una t, o el método de los intervalos de confianza. Si se usa la t, se tiene que

$$t = \frac{\bar{y}_{1.} - \bar{y}_{5.}}{s\sqrt{2/n}},$$

donde s es la raíz cuadrada del cuadrado medio del error y $n = 6$ es el tamaño de la muestra por tratamiento. En este caso,

$$t = \frac{553.33 - 610.67}{\sqrt{4961}\sqrt{1/3}} = -1.41.$$

El valor P para la prueba t con 25 grados de libertad es 0.17. Así que no hay evidencia suficiente para rechazar H_0.

Relación entre T y F

Anteriormente se expuso el uso de una prueba t agrupada, junto con los lineamientos que se estudiaron en el capítulo 10. El estimado agrupado se tomó del cuadrado medio del error con el fin de aprovechar los grados de libertad que están agrupados en las cinco muestras. Además, probamos un contraste. El lector debería observar que si el valor t se eleva al cuadrado, el resultado tiene exactamente la misma forma que el valor de f para una prueba del contraste, analizada en la sección anterior. En efecto,

$$f = \frac{(\bar{y}_{1.} - \bar{y}_{5.})^2}{s^2(1/6 + 1/6)} = \frac{(553.33 - 610.67)^2}{4961(1/3)} = 1.988,$$

que es, por supuesto, t^2.

Método del intervalo de confianza para una comparación por pares

Es fácil resolver el mismo problema de una comparación por pares (o un contraste) usando el método del intervalo de confianza. Es claro que, si se calcula un intervalo de confianza del $100(1-\alpha)\%$ sobre $\mu_1 - \mu_5$, se tiene que

$$\bar{y}_{1.} - \bar{y}_{5.} \pm t_{\alpha/2} s \sqrt{\frac{2}{6}},$$

donde $t_{\alpha/2}$ es el punto superior de $100(1-\alpha/2)\%$ de una distribución t con 25 grados de libertad (grados de libertad que provienen de s^2). Esta conexión inmediata entre las pruebas de hipótesis y los intervalos de confianza debería ser evidente a partir de los análisis de los capítulos 9 y 10. La prueba de un contraste simple $\mu_1 - \mu_5$ no implica más que observar si el intervalo de confianza anterior cubre o no al cero. Al sustituir los números se tiene que el intervalo de confianza de 95%:

$$(553.33 - 610.67) \pm 2.060\sqrt{4961}\sqrt{\frac{1}{3}} = -57.34 \pm 83.77.$$

Por consiguiente, como el intervalo de confianza cubre al cero, el contraste no es significativo. En otras palabras, no hay diferencia significativa entre las medias de los agregados 1 y 5.

Tasa de error por experimento

Se presentan muchas dificultades cuando el analista intenta hacer muchas o todas las comparaciones por pares posibles. Para el caso de k medias habrá, desde luego, $r = k(k-1)/2$ comparaciones por pares posibles. Si se suponen comparaciones independientes, la tasa de error por experimento o tasa de error por familia, es decir, la probabilidad de un falso rechazo de al menos una de las hipótesis, es dada por $1-(1-\alpha)^r$, donde α es la probabilidad seleccionada del error tipo I para una comparación específica. Es claro que esta medida del error tipo I por experimento sería bastante grande. Por

ejemplo, aun si sólo hubiera 6 comparaciones, digamos, en el caso de 4 medias, y $\alpha = 0.05$, la tasa de experimento-juicio sería

$$1 - (0.95)^6 \approx 0.26.$$

Cuando se prueban muchas comparaciones por pares, por lo general existe la necesidad de hacer el contraste efectivo sobre una sola comparación más conservadora. Es decir, usando el método del intervalo de confianza, los intervalos de confianza serían mucho más anchos que $\pm t_{\alpha/2} s\sqrt{2/n}$ que se emplea para el caso de una sola comparación.

Prueba de Tukey

Hay varios métodos estándar para realizar comparaciones por pares que den credibilidad a la tasa del error tipo I. Aquí se analizarán e ilustrarán dos de ellos. El primero, denominado **procedimiento de Tukey**, permite la formación de intervalos de confianza del $100(1-\alpha)\%$ simultáneos para todas las comparaciones por pares. El método se basa en la distribución del rango *estudentizado*. El punto apropiado del percentil es una función de α, k y v = grados de libertad para s^2. En la tabla A.12 se presenta una lista de puntos porcentuales superiores adecuados para $\alpha = 0.05$. El método de Tukey de comparaciones por pares implica encontrar una diferencia significativa entre las medias i y j $(i \neq j)$ si $|\bar{y}_{i.} - \bar{y}_{j.}|$ excede a $q(\alpha, k, v\sqrt{\frac{s^2}{n}}$.

El procedimiento de Tukey se ilustra con facilidad. Considere un ejemplo hipotético en el que se tienen 6 tratamientos en un diseño completamente aleatorizado de un solo factor, en el que se hacen 5 observaciones por tratamiento. Suponga que el cuadrado medio del error tomado de la tabla del análisis de varianza es $s^2 = 2.45$ (24 grados de libertad). Las medias muestrales están en orden ascendente,

$\bar{y}_{2.}$	$\bar{y}_{5.}$	$\bar{y}_{1.}$	$\bar{y}_{3.}$	$\bar{y}_{6.}$	$\bar{y}_{4.}$
14.50	16.75	19.84	21.12	22.90	23.20.

Con $\alpha = 0.05$, el valor de $q(0.05, 6, 24)$ es 4.37. Así, todas las diferencias absolutas se comparan con

$$4.37\sqrt{\frac{2.45}{5}} = 3.059.$$

Como resultado, las siguientes representan medias que, usando el procedimiento de Tukey, se encuentra que son significativamente diferentes:

4 y 1,	4 y 5,	4 y 2,	6 y 1,	6 y 5,
6 y 2,	3 y 5,	3 y 2,	1 y 5,	1 y 2.

¿De dónde proviene el nivel α en la prueba de Tukey?

Se mencionó brevemente el concepto de **intervalos de confianza simultáneos** que se emplean para el procedimiento de Tukey. El lector obtendrá una perspectiva útil del concepto de comparaciones múltiples, si comprende el significado de los intervalos de confianza simultáneos.

En el capítulo 9 vimos que, si se calcula un intervalo de confianza de 95% para, digamos, una media μ, entonces la probabilidad de que el intervalo cubra la media verdadera μ es 0.95.

Sin embargo, como vimos antes, para el caso de comparaciones múltiples la probabilidad efectiva de interés está ligada con la tasa de error por experimento, y debe hacerse énfasis en que los intervalos de confianza del tipo $\bar{y}_{i.} - \bar{y}_{j.} \pm q(\alpha, k, v)s\sqrt{1/n}$ no son independientes, ya que todos implican a s y muchos utilizan los mismos promedios, las $\bar{y}_{i.}$. A pesar de tales dificultades, si se utiliza la $q(0.05, k, v)$, el nivel de confianza simultáneo está controlado en un 95%. Lo mismo es cierto para $q(0.01, k, v)$, es decir, el nivel de confianza está controlado en un 99%. En el caso de $\alpha = 0.05$, hay una probabilidad de 0.05 de que se encuentre falsamente que al menos un par de mediciones son diferentes (falso rechazo de al menos una hipótesis nula). En el caso de $\alpha = 0.01$, la probabilidad correspondiente será 0.01.

13.9. Comparación de un conjunto de tratamientos en bloques

En la sección 13.2 estudiamos la idea de la formación de bloques, es decir, de aislar conjuntos de unidades experimentales que son razonablemente homogéneas y asignarles tratamientos de forma aleatoria. Ésta es una extensión del concepto de "formar pares" que se analizó en los capítulos 9 y 10, y se hace para reducir el error experimental, ya que las unidades en un bloque tienen más características comunes que las unidades localizadas en diferentes bloques.

El lector no debería considerar los bloques como un segundo factor, aunque ésa sea una forma tentadora de visualizar el diseño. De hecho, el factor principal (los tratamientos) aún lleva el peso mayor del experimento. Las unidades experimentales siguen siendo la fuente del error, igual que en el diseño completamente aleatorizado. Con la formación de bloques simplemente tratamos a dichas unidades de manera más sistemática. De ese modo, se dice que la aleatoriedad tiene restricciones. Antes de iniciar el estudio de la formación de bloques revisaremos dos ejemplos de un **diseño completamente aleatorizado**. El primer ejemplo es un experimento químico diseñado para determinar si hay una diferencia en la reacción media producida por cuatro catalizadores. Las muestras de los materiales que tienen que probarse se extraen de los mismos lotes de materias primas, a la vez que se mantienen constantes otras condiciones como la temperatura y concentración de los reactivos. En este caso, la hora del día en que se efectúan las corridas experimentales podría representar las unidades experimentales, y si el experimentador considera que es posible que haya un ligero efecto del tiempo, aleatorizaría la asignación

de los catalizadores a las corridas para contrarrestar la posible tendencia. Como un segundo ejemplo de dicho diseño, considere un experimento para comparar cuatro métodos para medir una propiedad física en particular de un fluido. Suponga que el proceso de muestreo es destructivo, es decir, que una vez que se ha medido una muestra de la sustancia usando un método, ya no puede medirse nuevamente con ningún otro. Si se decide hacer cinco mediciones con cada método, entonces se seleccionan al azar 20 muestras del material de un lote grande y se utilizan en el experimento para comparar los cuatro métodos de medición. Las unidades experimentales son las muestras seleccionadas al azar. Cualquier variación de una muestra a otra aparecerá en la variación del error, según se mida con s^2 en el análisis.

¿Cuál es el propósito de formar bloques?

Si la variación debida a la heterogeneidad en las unidades experimentales es tan grande que la sensibilidad para detectar diferencias de tratamiento se reduce debido a un valor aumentado de s^2, un plan mejor sería "bloquear" la variación debida a esas unidades y, por consiguiente, reducir la variación ajena a la que es explicada por bloques más pequeños o más homogéneos. Por ejemplo, suponga que en el ejemplo anterior de los catalizadores se supiera *a priori* que existe en definitiva un efecto significativo diario sobre el producto, y que es posible medir el producto para cuatro catalizadores en un día específico. En lugar de asignar los 4 catalizadores a las 20 corridas de prueba completamente al azar, se eligen, por ejemplo, 5 días y se prueba cada uno de los cuatro catalizadores cada día, asignándolos al azar a las corridas dentro de los días. De esta manera se elimina la variación diaria del análisis y, en consecuencia, el error experimental, que aún incluye cualquier tendencia temporal *dentro de los días*, representa con más precisión la variación aleatoria. A cada día se le denomina **bloque**.

El más directo de los diseños aleatorizados de bloques es aquel en el cual se asigna al azar un tratamiento a la vez a cada bloque. A un plan experimental como éste se le denomina **diseño de bloques completos aleatorizados (BCA)** y cada bloque constituye una sola réplica de los tratamientos.

13.10. Diseño de bloques completos aleatorios

Un plan clásico para el diseño de bloques completos aleatorizados (BCA) usando tres mediciones en cuatro bloques es el siguiente:

Bloque 1	Bloque 2	Bloque 3	Bloque 4
t_2	t_1	t_3	t_2
t_1	t_3	t_2	t_1
t_3	t_2	t_1	t_3

Las t denotan la asignación de cada uno de 3 tratamientos a los bloques. Desde luego, la asignación verdadera de los tratamientos a las unidades dentro de los bloques se hace al azar. Una vez que ha finalizado el experimento, los datos se pueden registrar como en el siguiente arreglo de 3×4:

Tratamiento	Bloque: 1	2	3	4
1	y_{11}	y_{12}	y_{13}	y_{14}
2	y_{21}	y_{22}	y_{23}	y_{24}
3	y_{31}	y_{32}	y_{33}	y_{34}

donde y_{11} representa la respuesta que se obtiene al utilizar el tratamiento 1 en el bloque 1, y_{12} es la respuesta que se obtiene al utilizar el tratamiento 1 en el bloque 2,..., y y_{34} es la respuesta que se obtiene al utilizar el tratamiento 3 en el bloque 4.

Ahora vamos a generalizar y a considerar el caso de k tratamientos asignados a b bloques. Los datos se pueden resumir tal como se observa en el arreglo rectangular de $k \times b$ de la tabla 13.7. Se supondrá que las y_{ij}, $i = 1, 2,..., k$ y $j = 1,2,..., b$, son valores de variables aleatorias independientes que tienen distribuciones normales con media μ_{ij} y varianza común σ^2.

Tabla 13.7: Arreglo de $k \times b$ para el diseño de BCA

	Bloque							
Tratamiento	**1**	**2**	...	***j***	...	***b***	**Total**	**Media**
1	y_{11}	y_{12}	...	y_{1j}	...	y_{1b}	$T_{1.}$	$\bar{y}_{1.}$
2	y_{21}	y_{22}	...	y_{2j}	...	y_{2b}	$T_{2.}$	$\bar{y}_{2.}$
⋮	⋮	⋮		⋮		⋮	⋮	⋮
i	y_{i1}	y_{i2}	...	y_{ij}	...	y_{ib}	$T_{i.}$	$\bar{y}_{i.}$
⋮	⋮	⋮		⋮		⋮	⋮	⋮
k	y_{k1}	y_{k2}	...	y_{kj}	...	y_{kb}	$T_{k.}$	$\bar{y}_{k.}$
Total	$T_{.1}$	$T_{.2}$	...	$T_{.j}$	...	$T_{.b}$	$T_{..}$	
Media	$\bar{y}_{.1}$	$\bar{y}_{.2}$	...	$\bar{y}_{.j}$	...	$\bar{y}_{.b}$		$\bar{y}_{..}$

Sea $\mu_{i.}$ el promedio (en lugar del total) de las b medias de la población para el i-ésimo tratamiento. Es decir,

$$\mu_{i.} = \frac{1}{b}\sum_{j=1}^{b}\mu_{ij}, \text{ para } i = 1, \ldots, k.$$

De manera similar, el promedio de las medias de la población para el j-ésimo bloque, $\mu_{.j}$, es definido por

$$\mu_{.j} = \frac{1}{k}\sum_{i=1}^{k}\mu_{ij}, \text{ para } j = 1, \ldots, b$$

y el promedio de las bk medias de la población, μ, es definido por

$$\mu = \frac{1}{bk}\sum_{i=1}^{k}\sum_{j=1}^{b}\mu_{ij}.$$

Para determinar si parte de la variación de nuestras observaciones se debe a diferencias entre los tratamientos, se considera la siguiente prueba:

Hipótesis de medias iguales de los tratamientos	H_0: $\mu_{1.} = \mu_{2.} = \cdots \mu_{k.} = \mu$, H_1: No todas las $\mu_{i.}$ son iguales.

13.11. Métodos gráficos y diagnósticos posteriores

En varios capítulos de este libro se hace referencia a procedimientos gráficos para mostrar datos y resultados analíticos. En los primeros capítulos se usaron gráficas de tallo y hojas y de caja y extensión como auxiliares visuales para resumir muestras. En el capítulo 10 se emplearon diagnósticos similares para entender mejor los datos de dos problemas de muestreo. En el capítulo 11 se introdujo el concepto de gráfica de residuales para detectar violaciones de las suposiciones estándar. En los últimos años gran parte de la atención dedicada al análisis de datos se ha centrado en los **métodos gráficos**. Al igual que en la regresión, el análisis de varianza se presta a la elaboración de gráficas que ayudan a resumir los datos y a detectar violaciones. Por ejemplo, una gráfica sencilla de las observaciones brutas alrededor de la media de cada tratamiento proporciona al analista una noción de la variabilidad entre las medias muestrales y dentro de las muestras. La figura 13.7 ilustra una de tales gráficas para los datos de agregados que se presentan en la tabla 13.1. A partir de la apariencia de la gráfica se obtiene incluso una idea de cuáles agregados (si los hubiera) destacan de los demás. Es evidente que el agregado 4 resalta del resto, y que los agregados 3 y 5 forman un grupo homogéneo, así como los agregados 1 y 2.

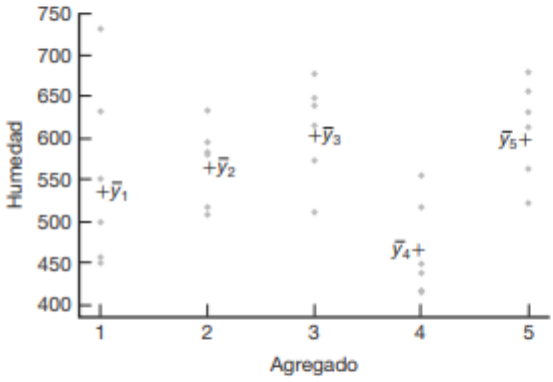


Figura 13.7: Gráfica de los datos alrededor de la media para los datos de los agregados de la tabla 13.1.

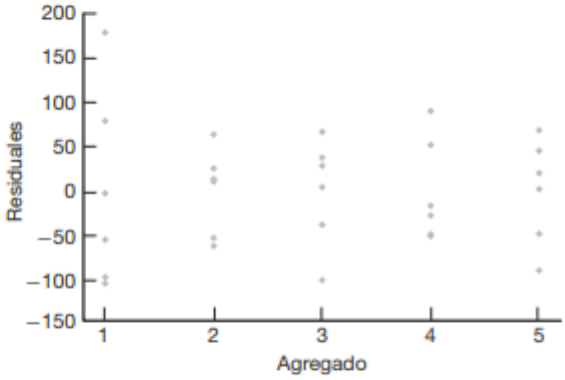


Figura 13.8: Gráfica de los residuales para cinco agregados con los datos de la tabla 13.1.

Como ocurre en el caso de la regresión, los residuales son útiles en el análisis de varianza para dar un diagnóstico sobre la detección de violaciones de los supuestos. Para formar los residuales sólo necesitamos considerar el modelo del problema de un solo factor, que es

$$y_{ij} = \mu_i + \epsilon_{ij} .$$

Es fácil determinar que el estimado de μ_i es $\bar{y}_i$. Por lo tanto, el ij-ésimo residual es $\bar{y}_{ij} - \bar{y}_i$, lo cual se extiende fácilmente al modelo de bloques completos aleatorizados. Sería aleccionador graficar los residuos para cada agregado con el fin de obtener cierta información sobre la suposición de varianza homogénea. Esta gráfica se muestra en la figura 13.8.

Las tendencias en gráficas como éstas podrían revelar dificultades en ciertas situaciones, especialmente cuando la violación de una suposición en particular se manifiesta en la gráfica. En el caso de la figura 13.8, los residuales parecen indicar que las varianzas *dentro de los tratamientos* son razonablemente homogéneas, excepto la del agregado 1. Hay cierta evidencia gráfica de que la varianza del agregado 1 es más grande que la del resto.

13.12. Transformaciones de datos en el análisis de varianza

En el capítulo 11 se puso mucha atención a la transformación de la respuesta y en situaciones para las que se ajustaba un modelo de regresión lineal a un conjunto de datos. Es evidente que se aplican los mismos conceptos a la regresión lineal múltiple, aunque esto no se analizó en el capítulo 12. En el estudio de los modelos de regresión se hizo énfasis en las transformaciones de y que producirían un modelo que se ajustara mejor a los datos que uno en el que la y ingresara de forma lineal. Por ejemplo, si la estructura del "tiempo" es de naturaleza exponencial, entonces una transformación logarítmica de

y linealiza la estructura y, por lo tanto, se anticipa más éxito cuando se utiliza la respuesta transformada.

Aunque el propósito fundamental de la transformación de datos que se ha analizado hasta este momento ha sido mejorar el ajuste del modelo, hay otras razones para transformar o reexpresar la respuesta y, y muchas de ellas se relacionan con las suposiciones que se hacen, por ejemplo, las suposiciones de las cuales depende la validez del análisis. Una suposición muy importante en el análisis de varianza es la de la varianza homogénea que se estudió antes en la sección 13.4. Se supone una **varianza común** σ^2. Si la varianza difiere mucho de un tratamiento a otro, y se realiza el ANOVA estándar que se estudia en este capítulo (y en otros posteriores), los resultados serían muy deficientes. En otras palabras, el análisis de varianza no es **robusto** respecto a la suposición de varianza homogénea. Como se ha dicho hasta el momento, se trata del motivo principal para la graficación de los residuales que estudiamos en la sección anterior y que ilustramos en las figuras 13.9, 13.10 y 13.11. Esas gráficas permiten detectar problemas debidos a una varianza no homogénea. Sin embargo, ¿qué hay que hacer al respecto? ¿Cómo se corrigen?

13.13. modelos de efectos aleatorios

A lo largo de este capítulo estudiamos los procedimientos del análisis de varianza en los que el objetivo principal es estudiar el efecto sobre ciertas respuestas de ciertos tratamientos fijos o predeterminados. Los experimentos en los que los tratamientos o los niveles de tratamiento son preseleccionados por el experimentador, y no elegidos al azar, se denominan **experimentos de efectos fijos**. Para el modelo de efectos fijos sólo se hacen inferencias acerca de los tratamientos específicos que se utilizaron en el experimento.

Con frecuencia es importante que el experimentador sea capaz de hacer inferencias acerca de una población de tratamientos a través de un experimento en el que los tratamientos empleados se elijan al azar de entre la población. Por ejemplo, un biólogo podría estar interesado en saber si hay o no una varianza significativa en alguna característica fisiológica debida a un tipo de animal. Los tipos de animales que en realidad se usan en el experimento se eligen al azar y representan los efectos del tratamiento. Un químico podría estar interesado en estudiar el efecto de los laboratorios sobre el análisis químico de

una sustancia; no le interesa un laboratorio en particular, sino una población grande de laboratorios. Así, podría seleccionar al azar un grupo de laboratorios y asignar muestras a cada uno para su análisis. Entonces, la inferencia estadística implicaría 1) probar si los laboratorios contribuyen o no a una varianza diferente de cero en los resultados de los análisis, y 2) estimar la varianza debida a los laboratorios y a la varianza dentro de los mismos.

Modelo y suposiciones para el modelo de efectos aleatorios

El **modelo de efectos aleatorios** de un solo factor se escribe como el modelo de efectos fijos, pero sus términos tienen significados diferentes. La respuesta $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$ es ahora un valor de la variable aleatoria

$$Y_{ij} = \mu + A_i + \epsilon_{ij}, \text{ con } i = 1, 2, \ldots, k \text{ y } j = 1, 2, \ldots, n,$$

donde las A_i tienen distribución normal e independiente con media igual a cero y varianza σ_α^2, y son independientes de las ϵ_{ij}. Al igual que para el modelo de efectos fijos, las ϵ_{ij} también tienen distribución normal e independiente con media igual a cero y varianza σ^2. Observe que para un experimento de efectos aleatorios, ya no se aplica la restricción de que $\sum_{i=1}^{k} \alpha_i = 0$.

Teorema 13.4: Para el modelo del análisis de varianza de efectos aleatorios de un solo factor,

$$E(SCT) = (k-1)\sigma^2 + n(k-1)\sigma_\alpha^2 \qquad \text{y} \qquad E(SCE) = k(n-1)\sigma^2.$$

La tabla 13.11 presenta los cuadrados medios esperados tanto para un experimento de efectos fijos como para uno de efectos aleatorios. Los cálculos para un experimento de efectos aleatorios se realizan exactamente de la misma forma que los del experimento de efectos fijos. Es decir, la suma de cuadrados, los grados de libertad y las columnas de los cuadrados medios en la tabla del análisis de varianza son iguales para ambos modelos.

Tabla 13.11: Cuadrados medios esperados para el experimento de un solo factor

Fuente de variación	Grados de libertad	Cuadrados medios	Cuadrados medios esperados	
			Efectos fijos	Efectos aleatorios
Tratamientos	$k-1$	s_1^2	$\sigma^2 + \frac{n}{k-1}\sum_i \alpha_i^2$	$\sigma^2 + n\sigma_\alpha^2$
Error	$k(n-1)$	s^2	σ^2	σ^2
Total	$nk-1$			

Para el modelo de efectos aleatorios la hipótesis de que todos los efectos del tratamiento son iguales a cero se escribe como sigue:

Hipótesis para un experimento de efectos aleatorios

$$H_0\colon \sigma_\alpha^2 = 0,$$
$$H_1\colon \sigma_\alpha^2 \neq 0.$$

Esta hipótesis afirma que los diferentes tratamientos no contribuyen en absoluto a la variabilidad de la respuesta. De la tabla 13.11 es evidente que tanto s_1^2 como s^2 son estimados de σ^2 cuando H_0 es verdadera, y que la razón

$$f = \frac{s_1^2}{s^2}$$

es un valor de la variable aleatoria F que tiene la distribución F con $k-1$ y $k(n-1)$ grados de libertad. La hipótesis nula se rechaza a un nivel de significancia α cuando

$$f > f_\alpha[k-1, k(n-1)].$$

En muchos estudios científicos y de ingeniería el interés no se centra en la prueba F. El científico sabe que el efecto aleatorio, en efecto, es significativo. Lo más importante es la estimación de los diversos componentes de la varianza. Esto produce un sentido de *jerarquía* en términos de cuáles factores producen la mayor variabilidad y en qué cantidad. En este contexto podría ser interesante cuantificar cuánto más grande es el *componente de la varianza de un solo factor* que el producido por el azar (variación aleatoria).

Estimación de los componentes de la varianza

La tabla 13.11 también se utiliza para estimar los **componentes de la varianza** σ^2 y σ_α^2. Como s_1^2 estima $\sigma^2 + n\sigma_\alpha^2$, y s^2 estima σ^2,

$$\hat{\sigma}^2 = s^2, \qquad \hat{\sigma}_\alpha^2 = \frac{s_1^2 - s^2}{n}.$$

13.14. Estudio de caso

Estudio de caso 13.1: **Análisis químico**. Se pidió al personal del Departamento de Química de Virginia Tech que analizara un conjunto de datos que se obtuvo para comparar 4 métodos distintos de análisis del aluminio en cierta mezcla deflagradora sólida. Para considerar una amplia gama de laboratorios de análisis se utilizaron 5 de ellos en el experimento. Se seleccionaron esos laboratorios porque suelen realizar esa clase de análisis. Se asignaron al azar 20 muestras de material deflagrador que contenían 2.70% de aluminio, cuatro a cada laboratorio, y se dieron instrucciones acerca de cómo efectuar los análisis químicos utilizando los cuatro métodos. Los datos que se obtuvieron son los siguientes:

	Laboratorio					
Método	**1**	**2**	**3**	**4**	**5**	**Media**
A	2.67	2.69	2.62	2.66	2.70	2.668
B	2.71	2.74	2.69	2.70	2.77	2.722
C	2.76	2.76	2.70	2.76	2.81	2.758
D	2.65	2.69	2.60	2.64	2.73	2.662

Los laboratorios no se consideran efectos aleatorios, ya que no fueron seleccionados al azar de entre una población más grande de ellos. Los datos se analizaron como un diseño de bloques completos aleatorizados. Se dibujaron gráficas de los datos para determinar si era apropiado un modelo aditivo del tipo:

$$y_{ij} = \mu + m_i + l_j + \epsilon_{ij}$$

en otras palabras, un modelo con efectos aditivos. El bloque aleatorizado no es adecuado cuando existe interacción entre los laboratorios y los métodos. Considere la gráfica de la figura 13.12. Aunque es un poco difícil de interpretar porque cada punto representa una sola observación, parece que no hay interacción evidente entre los métodos y los laboratorios.

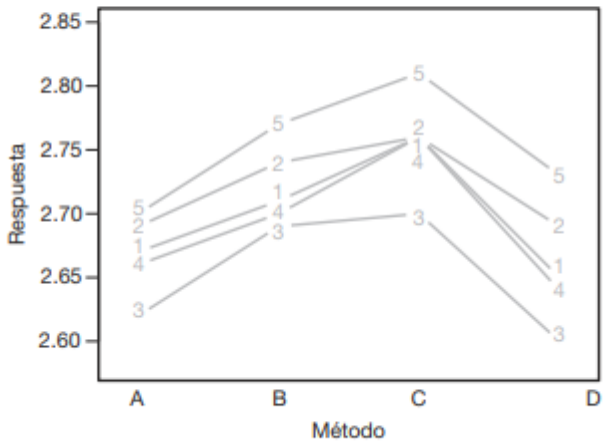


Figura 13.12: Gráfica de interacción para los datos del estudio de caso 13.1.

Gráficas de residuales

Las gráficas de residuales se usaron como indicaciones de diagnóstico con respecto a la suposición de una varianza homogénea. La figura 13.13 presenta una gráfica de residuales contra los métodos de análisis. La variabilidad descrita en los residuales parece ser bastante homogénea. Para completar, en la figura 13.14 se presenta una gráfica de probabilidad normal de los residuales.

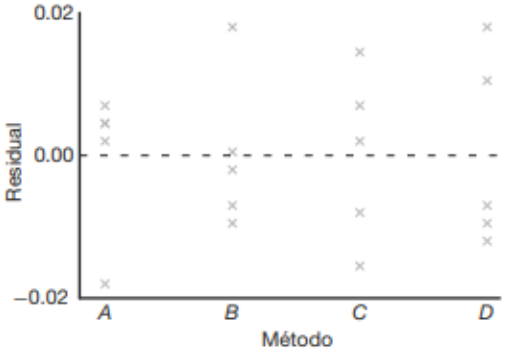


Figura 13.13: Gráfica de residuales en comparación con el método para los datos del estudio de caso 13.1.

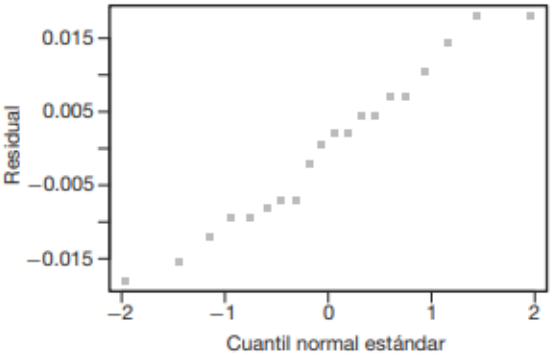


Figura 13.14: Gráfica de probabilidad normal de residuales para los datos del estudio de caso 13.1.

Las gráficas de residuales no muestran problemas con la suposición de errores normales ni con la de varianza homogénea. Para hacer el análisis de varianza se utilizó la función PROC GLM del programa *SAS*. En la figura 13.15 se incluye una salida de resultados por computadora con comentarios.

Los valores f y P calculados sí indican una diferencia significativa entre los métodos de análisis. A este análisis le puede seguir un análisis de comparación múltiple para determinar en dónde están las diferencias entre los métodos.

14. EXPERIMENTOS FACTORIALES

14.2.introducción

Considere una situación en la que haya interés por estudiar el efecto de **dos factores**, *A* y *B*, sobre alguna respuesta. Por ejemplo, en un experimento químico nos gustaría variar en forma simultánea la presión de reacción y el tiempo de reacción, y estudiar el efecto que cada uno tiene sobre el producto. En un experimento biológico resulta de interés estudiar el efecto que tienen el tiempo de secado y la temperatura sobre la cantidad de sólidos (porcentaje por peso) que queda en muestras de levadura. Igual que en el capítulo 13, el término **factor** se utiliza en un sentido general para denotar cualquier característica del experimento que pueda variar de un ensayo a otro, como la temperatura, el tiempo o la presión. Los **niveles** de un factor se definen como los valores reales que se utilizan en el experimento.

Para cada uno de estos casos es importante determinar no sólo si cada uno de los 2 factores influye en la respuesta, sino también si hay una interacción significativa entre ellos. En lo que se refiere a la terminología, el experimento descrito aquí es de 2 factores, y el diseño experimental podría ser uno completamente aleatorizado, en el que las distintas combinaciones de tratamiento se asignan al azar a todas las unidades experimentales, o bien, un diseño de bloques completos aleatorizados, donde las combinaciones de factores se asignan al azar dentro de los bloques. En el ejemplo de la levadura, si se empleara un diseño completamente aleatorizado, las distintas combinaciones de tratamientos de temperatura y tiempo de secado se asignarían al azar a las muestras de levadura.

Muchos de los conceptos que se estudiaron en el capítulo 13 se extienden a 2 y 3 factores en este capítulo. El objetivo principal de este material es el uso del diseño completamente aleatorizado con un *experimento factorial*. Un experimento factorial con 2 factores implica ensayos experimentales (o uno solo) con todas las combinaciones de factores. Por ejemplo, en el caso de la temperatura y tiempo de secado con, digamos, 3 niveles de cada uno y $n = 2$ corridas por cada una de las 9 combinaciones, tendríamos un *experimento factorial de 2 factores en un diseño completamente aleatorizado*. Ninguno de ellos es un factor de bloqueo; nos interesa la manera en que cada uno influye en el porcentaje de sólidos en las muestras, y si interactúan o no. El biólogo dispondría de 18 muestras físicas de material que constituyen unidades experimentales. Luego, éstas se asignarían al azar a las 18 combinaciones (9 combinaciones de tratamiento, cada una de ellas por duplicado).

Antes de entrar en detalles analíticos, sumas de cuadrados y demás, sería interesante que el lector observe la clara conexión que existe entre lo que hemos descrito y la situación con el problema de un solo factor. Considere el experimento de la levadura. La explicación de los grados de libertad ayuda a que el lector o el analista visualicen la extensión. En un inicio, las 9 combinaciones de tratamientos deberían considerarse como si representaran un factor con 9 niveles (8 grados de libertad). Así, un vistazo inicial a los grados de libertad arroja lo siguiente:

Combinaciones de tratamiento	8
Error	9
Total	17

14.3. interacción y el experimento de dos factores

En el modelo de bloques aleatorizados que se estudió previamente se supuso que en cada bloque se toma una observación de cada tratamiento. Si la suposición del modelo es correcta, es decir, si los bloques y los tratamientos son los únicos efectos reales y no hay interacción, el valor esperado del cuadrado medio del error es la varianza del error experimental σ^2. Sin embargo, suponga que existe interacción entre los tratamientos y los bloques, como lo indica el modelo

$$y_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ij}$$

de la sección 13.8. El valor esperado del cuadrado medio del error entonces es dado por

$$E\left[\frac{SCE}{(b-1)(k-1)}\right] = \sigma^2 + \frac{1}{(b-1)(k-1)}\sum_{i=1}^{k}\sum_{j=1}^{b}(\alpha\beta)_{ij}^2 .$$

Los efectos del tratamiento y los bloques no aparecen en el cuadrado medio del error esperado, pero los efectos de la interacción sí. Entonces, si en el modelo hay interacción, el cuadrado medio del error refleja variación debida al error experimental más una contribución de la interacción y, para este plan experimental, no hay forma de separarlos.

La interacción y la interpretación de los efectos principales

Desde el punto de vista del experimentador, parecería necesario llegar a una prueba significativa sobre la existencia de una interacción, al separar la variación del error verdadero de aquel que se debe a la interacción. Los efectos principales, A y B, adoptan un significado distinto en presencia de la interacción. En el ejemplo biológico anterior el efecto que tiene el tiempo de secado sobre la cantidad de sólidos que quedan en la levadura muy bien podría depender de la temperatura a la que se expusieron las muestras. En general, podrían existir situaciones experimentales en las que el factor A tuviera un efecto positivo sobre la respuesta en un nivel del factor B; en tanto que con un nivel distinto de B el efecto de A sería negativo. Aquí se usa el término **efecto positivo** para indicar que el producto o la respuesta se incrementan conforme los niveles de un factor dado aumentan de acuerdo con cierto orden definido. En el mismo sentido, un **efecto negativo** corresponde a una disminución de la respuesta al aumentar los niveles del factor.

Considere, por ejemplo, los siguientes datos de temperatura (factor A con niveles t_1, t_2 y t_3 (en orden creciente) y tiempo de secado d_1, d_2 y d_3 (también en orden creciente). La respuesta es el porcentaje de sólidos. Estos datos son completamente hipotéticos y se dan para ilustrar un aspecto.

	B			
A	d_1	d_2	d_3	**Total**
t_1	4.4	8.8	5.2	18.4
t_2	7.5	8.5	2.4	18.4
t_3	9.7	7.9	0.8	18.4
Total	21.6	25.2	8.4	55.2

Es evidente que el efecto de la temperatura sobre el porcentaje de sólidos es positivo para el tiempo breve de secado d_1, pero negativo para el tiempo prolongado d_3. Esta **interacción clara** entre la temperatura y el tiempo de secado es evidentemente interesante para el biólogo; sin embargo, con base en los totales de las respuestas para las temperaturas t_1, t_2 y t_3, la suma de cuadrados de la temperatura, *SCT*, producirá un valor de 0. Entonces, se dice que la presencia de la interacción **enmascara** el efecto de la temperatura. Por ello, si se considera el efecto medio de la temperatura, promediado para el tiempo de secado, **no existe efecto alguno**. Entonces, esto define el efecto principal. Pero, por supuesto, es probable que esto no sea pertinente para el biólogo.

Antes de sacar cualquier conclusión final a partir de las pruebas de significancia sobre los efectos principales y los efectos de la interacción, el **experimentador debería observar primero si la prueba para la interacción es o no significativa**. Si la interacción no es significativa, entonces los resultados de las pruebas sobre los efectos principales son importantes. No obstante, si la interacción debe ser significativa, entonces solamente son importantes aquellas pruebas sobre los efectos principales que resultan significativas. En presencia de una interacción, los efectos principales no significativos bien podrían ser resultado de enmascaramiento e indicar la necesidad de observar la influencia de cada factor a niveles fijos del otro.

Representación gráfica de la interacción

La presencia de interacción, así como su impacto científico, se puede interpretar adecuadamente usando **gráficas de interacción**. Las gráficas proporcionan una clara imagen de la tendencia de los datos para mostrar el efecto que tiene el cambio de un factor conforme se pasa de un nivel a otro del segundo factor. La figura 14.1 ilustra la fuerte interacción entre la temperatura y el tiempo de secado. La interacción se revela en las líneas no paralelas.

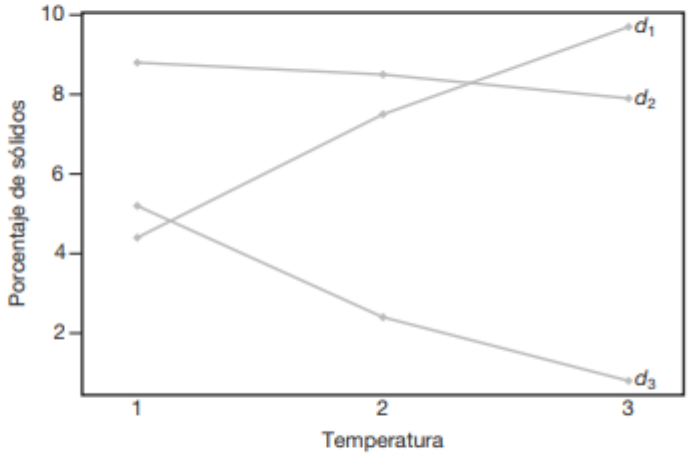


Figura 14.1: Gráfica de la interacción para los datos de temperatura y de tiempo de secado.

El *efecto relativamente fuerte de la temperatura* sobre el porcentaje de sólidos en el tiempo de secado más breve se refleja en la marcada pendiente de d_1. En el tiempo de secado medio, d_2, la temperatura tiene muy poco efecto, mientras que en el tiempo de secado prolongado d_3 la pendiente negativa indica un efecto negativo de la temperatura. Las gráficas de interacción como ésta le permiten al científico hacer una interpretación rápida y significativa de la interacción que existe. Debe ser evidente que el **paralelismo** en las gráficas indica la **ausencia de interacción**.

14.4. análisis de varianza de dos factores

Al presentar las fórmulas generales para el análisis de varianza de un experimento de 2 factores utilizando observaciones repetidas en un diseño completamente aleatorizado, debe considerarse el caso de n réplicas de las combinaciones del tratamiento, determinadas por a niveles del factor A y b niveles del factor B. Las observaciones se podrían clasificar usando un arreglo rectangular, donde los renglones representan los niveles del factor A y las columnas representan los niveles del factor B. Cada combinación de tratamiento define una celda del arreglo. Así, se tienen ab celdas, cada una de las cuales contiene n observaciones. Se denota con y_{ijk}, la k-ésima observación tomada en el i-ésimo nivel del factor A y el j-ésimo nivel del factor B. En la tabla 14.1 se muestran las abn observaciones.

Tabla 14.1: Experimento de dos factores con n réplicas

	B					
A	1	2	$\cdots$	b	Total	Media
1	y_{111}	y_{121}	$\cdots$	y_{1b1}	$Y_{1..}$	$\bar{y}_{1..}$
	y_{112}	y_{122}	$\cdots$	y_{1b2}		
	$\vdots$	$\vdots$		$\vdots$		
	y_{11n}	y_{12n}	$\cdots$	y_{1bn}		
2	y_{211}	y_{221}	$\cdots$	y_{2b1}	$Y_{2..}$	$\bar{y}_{2..}$
	y_{212}	y_{222}	$\cdots$	y_{2b2}		
	$\vdots$	$\vdots$		$\vdots$		
	y_{21n}	y_{22n}	$\cdots$	y_{2bn}		
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
a	y_{a11}	y_{a21}	$\cdots$	y_{ab1}	$Y_{a..}$	$\bar{y}_{a..}$
	y_{a12}	y_{a22}	$\cdots$	y_{ab2}		
	$\vdots$	$\vdots$		$\vdots$		
	y_{a1n}	y_{a2n}	$\cdots$	y_{abn}		
Total	$Y_{.1.}$	$Y_{.2.}$	$\cdots$	$Y_{.b.}$	$Y_{...}$	
Media	$\bar{y}_{.1.}$	$\bar{y}_{.2.}$	$\cdots$	$\bar{y}_{.b.}$		$\bar{y}_{...}$

definen los siguientes símbolos útiles, algunos de los cuales se utilizan en la tabla 14.1:

$Y_{ij.}$ = suma de las observaciones en la (ij)-ésima celda,
$Y_{i..}$ = suma de las observaciones para el i-ésimo nivel del factor A,
$Y_{.j.}$ = suma de las observaciones para el j-ésimo nivel del factor B,
$Y_{...}$ = suma de todas las abn observaciones,
$\bar{y}_{ij.}$ = media de las observaciones en la (ij)-ésima celda,
$\bar{y}_{i..}$ = media de las observaciones para el i-ésimo nivel del factor A,
$\bar{y}_{.j.}$ = media de las observaciones para el j-ésimo nivel del factor B,
$\bar{y}_{...}$ = media de todas las abn observaciones.

A diferencia de la situación para un solo factor, que se cubrió con amplitud en el capítulo 13, en éste supondremos que las **poblaciones**, de las que se toman n observaciones independientes con distribución idéntica, son **combinaciones** de los factores. Asimismo, se supondrá siempre que de cada combinación de factores se toma un número igual (n) de observaciones. En los casos en que los tamaños de las muestras por combinación son desiguales, los cálculos son más complicados, aunque los conceptos son transferibles.

Modelo e hipótesis para el problema de dos factores

Cada observación de la tabla 14.1 se puede escribir en la siguiente forma:

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk},$$

donde ϵ_{ijk} mide las desviaciones de los valores y_{ijk} observados en la (ij)-ésima celda a partir de la media de la población μ_{ij}. Si $(\alpha\beta)_{ij}$ denota el efecto de la interacción del i-ésimo nivel del factor A y el j-ésimo nivel del factor B, α_i el efecto del i-ésimo nivel del factor A, β_j el efecto del j-ésimo nivel del factor B, y μ la media conjunta, escribimos

$$\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij},$$

y, entonces,

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk},$$

a las que se imponen las restricciones

$$\sum_{i=1}^{a} \alpha_i = 0, \qquad \sum_{j=1}^{b} \beta_j = 0, \qquad \sum_{i=1}^{a} (\alpha\beta)_{ij} = 0, \qquad \sum_{j=1}^{b} (\alpha\beta)_{ij} = 0.$$

Las 3 hipótesis por probar son las siguientes:

1. H_0': $\alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$,
H_1': Al menos una de las α_i no es igual a 0.

2. H_0'': $\beta_1 = \beta_2 = \cdots = \beta_b = 0$,
H_1'': Al menos una de las β_j no es igual a 0.

3. H_0''': $(\alpha\beta)_{11} = (\alpha\beta)_{12} = \cdots = (\alpha\beta)_{ab} = 0$,

H_1''': Al menos una de las $(\alpha\beta)_{ij}$ no es igual a 0.

Se alerta al lector acerca del problema del enmascaramiento de los efectos principales cuando la interacción contribuye de manera importante en el modelo. Se recomienda considerar primero el resultado de la prueba de interacción y, luego, la interpretación de la prueba del efecto principal; la naturaleza de la conclusión científica depende de si se encontró interacción. Si ésta se descarta, entonces se pueden probar las hipótesis 1 y 2 y la interpretación es muy sencilla. Sin embargo, si se descubre que hay interacción, la interpretación puede ser más complicada, como se vio al analizar el tiempo de secado y la temperatura en la sección previa. La estructura de las pruebas de hipótesis 1, 2 y 3 se estudiará en las secciones siguientes. En el análisis del ejemplo 14.1 se incluirá la interpretación de los resultados.

Las pruebas de las hipótesis anteriores se basarán en la comparación de estimados independientes de σ^2, obtenidos al separar la suma de cuadrados total de los datos en 4 componentes mediante la siguiente identidad.

Partición de la variabilidad en el caso de dos factores

Teorema 14.1: Identidad de la suma de cuadrados

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{...})^2 = bn\sum_{i=1}^{a}(\bar{y}_{i..} - \bar{y}_{...})^2 + an\sum_{j=1}^{b}(\bar{y}_{.j.} - \bar{y}_{...})^2$$
$$+ n\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{ij}$$

Simbólicamente, la identidad de la suma de cuadrados se escribe como

$$SCT = SCA + SCB + SC(AB) + SCE$$

donde a SCA y SCB se les denomina la suma de cuadrados para los efectos principales A y B, respectivamente, $SC(AB)$ recibe el nombre de suma de cuadrados de la interacción para A y B, y SCE es la suma de cuadrados del error. La partición de los grados de libertad se efectúa de acuerdo con la identidad

$$abn - 1 = (a-1) + (b-1) + (a-1)(b-1) + ab(n-1).$$

Formación de los cuadrados medios

Si dividimos cada una de las sumas de cuadrados en el lado derecho de la identidad de la suma de cuadrados entre su número correspondiente de grados de libertad, obtenemos los cuatro estadísticos

$$S_1^2 = \frac{SCA}{a-1}, \qquad S_2^2 = \frac{SCB}{b-1}, \qquad S_3^2 = \frac{SC(AB)}{(a-1)(b-1)}, \qquad S^2 = \frac{SCE}{ab(n-1)}.$$

Todos estos estimados de la varianza son estimados independientes de σ^2, siempre que no haya efectos α_i, β_j ni, por supuesto, $(\alpha\beta)_{ij}$. Si las sumas de cuadrados se interpretan

como funciones de las variables aleatorias independientes $y_{111}, y_{112}, \ldots, y_{abn}$, no es difícil comprobar que

$$E(S_1^2) = E\left[\frac{SCA}{a-1}\right] = \sigma^2 + \frac{nb}{a-1}\sum_{i=1}^{a}\alpha_i^2,$$

$$E(S_2^2) = E\left[\frac{SCB}{b-1}\right] = \sigma^2 + \frac{na}{b-1}\sum_{j=1}^{b}\beta_j^2,$$

$$E(S_3^2) = E\left[\frac{SC(AB)}{(a-1)(b-1)}\right] = \sigma^2 + \frac{n}{(a-1)(b-1)}\sum_{i=1}^{a}\sum_{j=1}^{b}(\alpha\beta)_{ij}^2,$$

$$E(S^2) = E\left[\frac{SCE}{ab(n-1)}\right] = \sigma^2,$$

a partir de lo cual se observa de inmediato que los 4 estimados de σ^2 son no sesgados cuando H_0', H_0'', y H_0''' son verdaderas.

Para probar la hipótesis H_0', de que los efectos de los factores A son todos iguales a cero, se calcula la siguiente razón:

Prueba F para el factor A

$$f_1 = \frac{s_1^2}{s^2},$$

que es un valor de la variable aleatoria F_1, el cual tiene la distribución F con $a-1$ y $ab(n-1)$ grados de libertad cuando H_0', es verdadera. La hipótesis nula se rechaza al nivel de significancia α cuando $f_1 > f_\alpha[a-1, ab(n-1)]$.

De manera similar, para probar la hipótesis H_0'', de que todos los efectos del factor B son iguales a cero, se calcula la razón:

Prueba F para el factor B

$$f_2 = \frac{s_2^2}{s^2},$$

que es un valor de la variable aleatoria F_2 que tiene la distribución F con $b-1$ y $ab(n-1)$ grados de libertad cuando H_0'', es verdadera. Esta hipótesis se rechaza al nivel de significancia α cuando $f_2 > f_\alpha[b-1, ab(n-1)]$.

Por último, para probar la hipótesis H_0''', de que todos los efectos de interacción son iguales a 0, se calcula la razón siguiente:

Prueba F para la interacción

$$f_3 = \frac{s_3^2}{s^2},$$

que es un valor de la variable aleatoria F_3, el cual tiene la distribución F con $(a-1)(b-1)$ y $ab(n-1)$ grados de libertad cuando H_0''' es verdadera. Concluimos que, a un nivel de significancia α, hay interacción cuando $f_\alpha[(a-1)(b-1), ab(n-1)]$.

Como se indicó en la sección 14.2, se recomienda interpretar la prueba para la interacción antes de tratar de hacer inferencias sobre los efectos principales. Si la interacción no es significativa, entonces hay evidencia de que las pruebas sobre los efectos principales son interpretables. El rechazo de la hipótesis 1 de la página 566 implica que las medias de la respuesta en los niveles del factor A difieren significativamente, mientras que

el rechazo de la hipótesis 2 implica una condición similar para las medias en los niveles del factor B. Sin embargo, una interacción significativa podría muy bien implicar que los datos se deberían analizar de una manera un poco diferente, **quizá observando el efecto del factor A en niveles fijos del factor B**, y así sucesivamente.

Los cálculos en un problema de análisis de varianza para un experimento de 2 factores con n réplicas suelen resumirse como se ilustra en la tabla 14.2.

Tabla 14.2: Análisis de varianza para el experimento de 2 factores con n réplicas

Fuente de variación	**Suma de cuadrados**	**Grados de libertad**	**Cuadrado medio**	**f calculada**
Efecto principal				
A	SCA	$a-1$	$s_1^2 = \frac{SCA}{a-1}$	$f_1 = \frac{s_1^2}{s^2}$
B	SCB	$b-1$	$s_2^2 = \frac{SCB}{b-1}$	$f_2 = \frac{s_2^2}{s^2}$
Interacciones de 2 factores				
AB	$SC(AB)$	$(a-1)(b-1)$	$s_3^2 = \frac{SC(AB)}{(a-1)(b-1)}$	$f_3 = \frac{s_3^2}{s^2}$
Error	SCE	$ab(n-1)$	$s^2 = \frac{SCE}{ab(n-1)}$	
Total	STC	$abn-1$		

14.5.Análisis gráfico en el problema de dos factores

Muchos de los mismos tipos de ilustraciones gráficas que se sugirió emplear en los problemas de un factor también se aplican en el caso de 2 factores. Las gráficas en 2 dimensiones de las medias de las celdas o de las medias de las combinaciones de tratamientos ofrecen información sobre la presencia de interacciones entre los 2 factores.

Además, una gráfica de los residuales contra los valores ajustados bien podría indicar si se cumple o no la suposición de la varianza homogénea. Por supuesto, es frecuente que una violación de la suposición de varianza homogénea implique un aumento en la varianza del error conforme *los números de la respuesta se vuelven más grandes*. Como resultado, esta gráfica podría resaltar la violación.

La figura 14.3 presenta la gráfica de las medias de las celdas para el caso del propulsor de los sistemas de misiles del ejemplo 14.1. Observe gráficamente (en este caso) cuánta falta de paralelismo hay. Note el aplanamiento de la parte de la figura que indica el efecto del propulsor para el sistema 3. Esto ilustra la interacción entre los factores. La figura 14.4 muestra la gráfica de los residuales contra los valores ajustados para los mismos datos. Al parecer no hay dificultades con la suposición de la varianza homogénea.

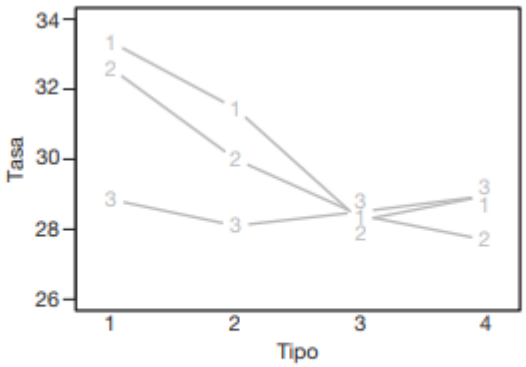


Figura 14.3: Gráfica de las medias de las celdas para los datos del ejemplo 14.1. Los números representan los sistemas de misiles.

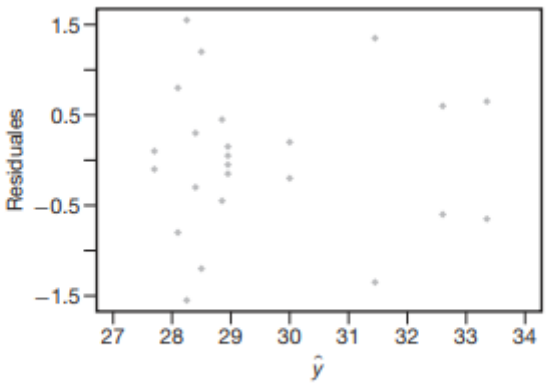


Figura 14.4: Gráfica de los residuales de los datos del ejemplo 14.1.

14.6.Experimentos de tres factores

En esta sección consideramos un experimento con 3 factores, A, B y C, en los niveles a, b y c, respectivamente, en un diseño experimental completamente aleatorizado. Suponga de nuevo que se tienen n observaciones para cada una de las abc combinaciones de tratamientos. Debemos proceder a realizar las pruebas de significancia para los 3 efectos principales y las interacciones implicadas. Se espera que el lector podrá utilizar después esta descripción para generalizar el análisis a $k > 3$ factores.

Modelo para el experimento de tres factores

El modelo para el experimento de 3 factores es

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkl},$$

$i = 1, 2, \ldots, a$; $j = 1, 2, \ldots, b$; $k = 1, 2, \ldots, c$; y $l = 1, 2, \ldots, n$, donde α_i, β_j y γ_k son los efectos principales y $(\alpha\beta)_{ij}$, $(\alpha\gamma)_{ik}$ y $(\beta\gamma)_{jk}$ son los efectos de la interacción de 2 factores que tienen la misma interpretación que en el experimento con 2 factores.

El término $(\alpha\beta\gamma)_{ijk}$ se denomina **efecto de interacción de 3 factores**, y representa la no aditividad de las $(\alpha\beta)_{ij}$ sobre los diferentes niveles del factor C. Igual que antes, la suma de todos los efectos principales es igual a 0, y la suma sobre cualesquiera de los subíndices de los efectos de la interacción entre 2 y 3 factores es igual a 0. En muchas situaciones experimentales estas interacciones de orden superior son insignificantes y sus cuadrados medios sólo reflejan variación aleatoria; pero se debe describir el análisis en su forma más general.

Nuevamente, para realizar pruebas válidas de significancia debe suponerse que los errores son valores de variables aleatorias independientes y con distribución normal, cada una con media igual a 0 y varianza común σ^2.

La filosofía general respecto al análisis es la misma que la que se estudió para los experimentos de 1 y 2 factores. La suma de cuadrados se divide en 8 términos, donde cada uno representa una fuente de variación de los que se obtienen estimados independientes de σ^2 cuando todos los efectos principales y de la interacción son iguales a 0. Si los efectos de cualquier factor dado o interacción no son iguales a 0, entonces el cuadrado medio estimará la varianza del error más un componente debido al efecto sistemático en cuestión.

Suma de cuadrados para un experimento de tres factores

$$SCA = bcn \sum_{i=1}^{a} (\bar{y}_{i...} - \bar{y}_{....})^2 \qquad SC(AB) = cn \sum_i \sum_j (\bar{y}_{ij..} - \bar{y}_{i...} - \bar{y}_{.j..} + \bar{y}_{....})^2$$

$$SCB = acn \sum_{j=1}^{b} (\bar{y}_{.j..} - \bar{y}_{....})^2 \qquad SC(AC) = bn \sum_i \sum_k (\bar{y}_{i.k.} - \bar{y}_{i...} - \bar{y}_{..k.} + \bar{y}_{....})^2$$

$$SCC = abn \sum_{k=1}^{c} (\bar{y}_{..k.} - \bar{y}_{....})^2 \qquad SC(BC) = an \sum_j \sum_k (\bar{y}_{.jk.} - \bar{y}_{.j..} - \bar{y}_{..k.} + \bar{y}_{....})^2$$

$$SC(ABC) = n \sum_i \sum_j \sum_k (\bar{y}_{ijk.} - \bar{y}_{ij..} - \bar{y}_{i.k.} - \bar{y}_{.jk.} + \bar{y}_{i...} + \bar{y}_{.j..} + \bar{y}_{..k.} - \bar{y}_{....})^2$$

$$STC = \sum_i \sum_j \sum_k \sum_l (y_{ijkl} - \bar{y}_{....})^2 \qquad SCE = \sum_i \sum_j \sum_k \sum_l (y_{ijkl} - \bar{y}_{ijk.})^2$$

Aunque en esta sección hacemos énfasis en la interpretación de una salida de resultados por computadora con comentarios, en vez de preocuparnos por cálculos laboriosos de sumas de cuadrados, ofrecemos lo siguiente como la suma de cuadrados para los 3 efectos principales y las interacciones. Observe la evidente extensión del problema de 2 factores a uno de 3.

Los promedios en las fórmulas se definen como sigue:

$\bar{y}_{...}$ = promedio de todas las $abcn$ observaciones,
$\bar{y}_{i...}$ = promedio de las observaciones para el i-ésimo nivel del factor A,
$\bar{y}_{.j.}$ = promedio de las observaciones para el j-ésimo nivel del factor B,
$\bar{y}_{..k}$ = promedio de las observaciones para el k-ésimo nivel del factor C,
$\bar{y}_{ij..}$ = promedio de las observaciones para el i-ésimo nivel de A y el j-ésimo nivel de B,
$\bar{y}_{i.k}$ = promedio de las observaciones para el i-ésimo nivel de A y el k-ésimo nivel de C,
$\bar{y}_{.jk}$ = promedio de las observaciones para el j-ésimo nivel de B y el k-ésimo nivel de C,
$\bar{y}_{ijk}$ = promedio de las observaciones para la (ijk)-ésima combinación de tratamientos.

Los cálculos en una tabla de análisis de varianza para un problema de 3 factores con n réplicas de corridas para cada combinación de factores se resumen en la tabla 14.7.

Tabla 14.7: ANOVA para el experimento de 3 factores con n réplicas

Fuente de variación	**Suma de cuadrados**	**Grados de libertad**	**Cuadrado medio**	***f* calculada**
Efecto principal:				
A	SCA	$a-1$	s_1^2	$f_1 = \frac{s_1^2}{s^2}$
B	SCB	$b-1$	s_2^2	$f_2 = \frac{s_2^2}{s^2}$
C	SCC	$c-1$	s_3^2	$f_3 = \frac{s_3^2}{s^2}$
Interacción de 2 factores:				
AB	$SC(AB)$	$(a-1)(b-1)$	s_4^2	$f_4 = \frac{s_4^2}{s^2}$
AC	$SC(AC)$	$(a-1)(c-1)$	s_5^2	$f_5 = \frac{s_5^2}{s^2}$
BC	$SC(BC)$	$(b-1)(c-1)$	s_6^2	$f_6 = \frac{s_6^2}{s^2}$
Interacción de 3 factores:				
ABC	$SC(ABC)$	$(a-1)(b-1)(c-1)$	s_7^2	$f_7 = \frac{s_7^2}{s^2}$
Error	SCE	$abc(n-1)$	s^2	
Total	STC	$abcn-1$		

Para el experimento de 3 factores con una sola corrida experimental por combinación se podría utilizar el análisis de la tabla 14.7 con $n = 1$ y usando la suma de cuadrados de la interacción ABC para SCE. En este caso suponemos que los efectos de la interacción $(\alpha\beta\gamma)_{ijk}$ son todos iguales a cero, de modo que

$$E\left[\frac{SC(ABC)}{(a-1)(b-1)(c-1)}\right] = \sigma^2 + \frac{n}{(a-1)(b-1)(c-1)}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(\alpha\beta\gamma)_{ijk}^2 = \sigma^2.$$

Es decir, $SC(ABC)$ representa la variación que sólo se debe al error experimental. Por lo tanto, su cuadrado medio proporciona un estimado no sesgado de la varianza del error. Con $n = 1$ y $SCE = SC(ABC)$, la suma de cuadrados del error se obtiene restando la suma de cuadrados de los efectos principales y las interacciones de 2 factores a la suma de cuadrados total.

Agrupamiento en modelos multifactoriales

El modelo de 3 factores y su análisis se describió de la manera más general mediante la inclusión en el modelo de todas las interacciones posibles. Por supuesto, hay muchas situaciones en las que *a priori* se sabe que el modelo no debería contener ciertas interacciones. Así, es posible aprovechar este conocimiento al combinar o agrupar las sumas de cuadrados correspondientes a interacciones despreciables con la suma de cuadrados del error para formar un nuevo estimador de σ^2 con un número más grande de grados de libertad. Por ejemplo, en un experimento de metalurgia diseñado para estudiar el efecto de 3 variables importantes del proceso sobre el espesor de película, suponga que se sabe que el factor A, la concentración de ácido, no interactúa con los factores B y C. Las

Tabla 14.11: ANOVA sin interacción del factor A

Fuente de variación	**Suma de cuadrados**	**Grados de libertad**	**Cuadrado medio**	***f* calculada**
Efecto principal:				
A	SCA	$a-1$	s_1^2	$f_1 = \frac{s_1^2}{s^2}$
B	SCB	$b-1$	s_2^2	$f_2 = \frac{s_2^2}{s^2}$
C	SCC	$c-1$	s_3^2	$f_3 = \frac{s_3^2}{s^2}$
Interacción de 2 factores:				
BC	$SC(BC)$	$(b-1)(c-1)$	s_4^2	$f_4 = \frac{s_4^2}{s^2}$
Error	SCE	Resta	s^2	
Total	STC	$abcn-1$		

sumas de cuadrados SCA, SCB, SCC y $SC(BC)$ se calculan usando los métodos descritos en un apartado anterior de esta sección. Todos los cuadrados medios de los efectos restantes ahora estimarán de manera independiente la varianza del error σ^2. Por lo tanto, formamos el nuevo **cuadrado medio del error agrupando** $SC(AB)$, $SC(AC)$, $SC(ABC)$ y SCE junto con los grados de libertad correspondientes. El denominador resultante de las pruebas de significancia es, entonces, el cuadrado medio del error dado por

$$s^2 = \frac{SC(AB) + SC(AC) + SC(ABC) + SCE}{(a-1)(b-1) + (a-1)(c-1) + (a-1)(b-1)(c-1) + abc(n-1)}.$$

Por supuesto, con una resta se obtienen la suma de cuadrados agrupada y los grados de libertad agrupados, una vez que se calcula la STC y las sumas de cuadrados para los efectos existentes. La tabla del análisis de varianza adoptaría así la forma de la tabla 14.11.

Experimentos factoriales en bloques

En este capítulo se ha supuesto que el diseño experimental utilizado es un diseño completamente aleatorizado. Al interpretar los niveles del factor A en la tabla 14.11 **como bloques diferentes** se tiene el procedimiento del análisis de varianza para un experimento de 2 factores en un diseño de bloques aleatorizados. Por ejemplo, si se interpretan los operadores del ejemplo 14.4 como bloques, y se supone que no hay interacción entre los bloques y los otros 2 factores, el análisis de varianza adopta la forma de la tabla 14.12, en vez de la de la tabla 14.9. El lector puede verificar que el cuadrado medio del error también es

$$s^2 = \frac{4.77 + 2.91 + 4.91 + 21.61}{4 + 2 + 4 + 36} = 0.74,$$

lo que demuestra el agrupamiento de las sumas de cuadrados para los efectos de la interacción inexistente. Observe que el factor B, el catalizador, tiene un efecto significativo sobre el producto.

Tabla 14.12: ANOVA para un experimento de 2 factores en un diseño de bloques aleatorizados

Fuente de variación	**Suma de cuadrados**	**Grados de libertad**	**Cuadrado medio**	***f* calculada**	**Valor *P***
Bloques	13.98	2	6.99		
Efecto principal:					
B	10.18	2	5.09	6.88	0.0024
C	1.18	1	1.18	1.59	0.2130
Interacción de 2 factores					
BC	3.64	2	1.82	2.46	0.0966
Error	34.21	46	0.74		
Total	63.19	53			

14.7.Experimentos factoriales II y III

En un experimento de 2 factores con efectos aleatorios se tiene el modelo

$$Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + \epsilon_{ijk},$$

para $i = 1, 2, \ldots, a$; $j = 1, 2, \ldots, b$; y $k = 1, 2, \ldots, n$, donde A_i, B_j $(AB)_{ij}$ y ϵ_{ijk} son variables aleatorias independientes con medias igual a 0 y varianzas σ^2_α, σ^2_β, $\sigma^2_{\alpha\beta}$ y σ^2, respectivamente. Las sumas de cuadrados para experimentos de efectos aleatorios se calculan

exactamente de la misma forma que en los experimentos de efectos fijos. Ahora se tiene interés en probar hipótesis con la forma

$$H_0^{'}\colon\ \sigma^2_\alpha = 0, \qquad H_0^{''}\colon\ \sigma^2_\beta = 0, \qquad H_0^{'''}\colon\ \sigma^2_{\alpha\beta} = 0,$$
$$H_1^{'}\colon\ \sigma^2_\alpha \neq 0, \qquad H_1^{''}\colon\ \sigma^2_\beta \neq 0, \qquad H_1^{'''}\colon\ \sigma^2_{\alpha\beta} \neq 0,$$

donde el denominador en la razón f no es necesariamente el cuadrado medio del error. El denominador apropiado se determina examinando los valores esperados de los distintos cuadrados medios, los cuales se presentan en la tabla 14.14.

Tabla 14.14: Cuadrados medios esperados para un experimento de efectos aleatorios de 2 factores

Fuente de variación	Grados de libertad	Cuadrado medio	Cuadrado medio esperado
A	$a-1$	s_1^2	$\sigma^2 + n\sigma^2_{\alpha\beta} + bn\sigma^2_\alpha$
B	$b-1$	s_2^2	$\sigma^2 + n\sigma^2_{\alpha\beta} + an\sigma^2_\beta$
AB	$(a-1)(b-1)$	s_3^2	$\sigma^2 + n\sigma^2_{\alpha\beta}$
Error	$ab(n-1)$	s^2	σ^2
Total	$abn-1$		

En la tabla 14.14 se observa que $H_0^{'}$ y $H_0^{''}$ se prueban usando s_3^2 en el denominador de la razón f; mientras que $H_0^{'''}$ se prueba con s^2 en el denominador. Los estimados no sesgados de los componentes de la varianza son

$$\hat{\sigma}^2 = s^2, \qquad \hat{\sigma}^2_{\alpha\beta} = \frac{s_3^2 - s^2}{n}, \qquad \hat{\sigma}^2_\alpha = \frac{s_1^2 - s_3^2}{bn}, \qquad \hat{\sigma}^2_\beta = \frac{s_2^2 - s_3^2}{an}.$$

Tabla 14.15: Cuadrados medios esperados para un experimento de efectos aleatorios de 3 factores

Fuente de variación	Grados de libertad	Cuadrado medio	Cuadrado medio esperado
A	$a-1$	s_1^2	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma} + cn\sigma^2_{\alpha\beta} + bn\sigma^2_{\alpha\gamma} + bcn\sigma^2_\alpha$
B	$b-1$	s_2^2	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma} + cn\sigma^2_{\alpha\beta} + an\sigma^2_{\beta\gamma} + acn\sigma^2_\beta$
C	$c-1$	s_3^2	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma} + bn\sigma^2_{\alpha\gamma} + an\sigma^2_{\beta\gamma} + abn\sigma^2_\gamma$
AB	$(a-1)(b-1)$	s_4^2	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma} + cn\sigma^2_{\alpha\beta}$
AC	$(a-1)(c-1)$	s_5^2	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma} + bn\sigma^2_{\alpha\gamma}$
BC	$(b-1)(c-1)$	s_6^2	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma} + an\sigma^2_{\beta\gamma}$
ABC	$(a-1)(b-1)(c-1)$	s_7^2	$\sigma^2 + n\sigma^2_{\alpha\beta\gamma}$
Error	$abc(n-1)$	s^2	σ^2
Total	$abcn-1$		

En la tabla 14.15 se presentan los cuadrados medios esperados para el experimento de 3 factores con efectos aleatorios en un diseño completamente aleatorizado. A partir de los cuadrados medios esperados de la tabla 14.15 es evidente que se pueden formar

razones f adecuadas para probar todos los componentes de la varianza de la interacción de 2 y 3 factores. Sin embargo, para probar una hipótesis de la forma

$$H_0: \ \sigma_\alpha^2 = 0,$$
$$H_1: \ \sigma_\alpha^2 \neq 0,$$

parece que no hay razón f apropiada, a menos que se encontrara que uno o más de los componentes de la varianza de interacción de 2 factores no es significativo. Por ejemplo, suponga que se hubiera comparado s_5^2 (cuadrado medio AC) con s_7^2 (cuadrado medio ABC) y se encontrara que $\sigma_{\alpha\gamma}^2$ es despreciable. Entonces podría argumentarse que el término $\sigma_{\alpha\gamma}^2$ debería eliminarse de todos los cuadrados medios esperados de la tabla 14.15; entonces, la razón s_1^2/s_4^2 ofrece una prueba de la significancia del componente σ_α^2 de la varianza. Por lo tanto, si se prueba la hipótesis concerniente a los componentes de la varianza de los efectos principales, es necesario investigar primero la significancia de los componentes de la interacción de 2 factores. Cuando se encuentra que ciertos componentes de la varianza de la interacción de 2 factores son significativos, por lo que deben permanecer como parte del cuadrado medio esperado, se utiliza una prueba aproximada derivada por Satterthwaite (1946; véase la bibliografía) .

14.8. Posibles riesgos y errores conceptuales; relación con el material de otros capítulos

Uno de los temas más susceptibles de confusión en el análisis de experimentos factoriales radica en la interpretación de los efectos principales ante la presencia de interacción. La existencia de un valor P relativamente grande para un efecto principal, cuando es clara la presencia de interacciones, podría tentar al analista a concluir que "no existe efecto principal significativo". Sin embargo, debe entenderse que si un efecto principal está implicado en una interacción significativa, entonces el efecto principal **está influyendo en la respuesta**. La naturaleza del efecto es inconsistente a través de los niveles de otros efectos. La naturaleza del papel que desempeña el efecto principal se deduce de las **gráficas de interacción**.

Debido a lo que se expresa en el párrafo anterior, hay un gran peligro de usar la estadística de manera equivocada cuando se emplea una prueba de comparación múltiple sobre los efectos principales ante la presencia clara de interacción entre los factores.

Debe tenerse precaución en el análisis de un experimento factorial cuando se supone un diseño completamente aleatorizado y en realidad no se hizo tal aleatorización. Por ejemplo, es común que se encuentren factores que son **muy difíciles de cambiar**. Como resultado, podría ser necesario mantener sin cambio los niveles de factores durante largos periodos a lo largo de todo el experimento. El ejemplo más común es el factor temperatura. Subirla o bajarla en un esquema aleatorio es un plan costoso y la mayoría de los experimentadores evitarán hacerlo. Los diseños experimentales con *restricciones en la aleatorización* son muy comunes y reciben el nombre de **diseños de gráficas separadas**. Esos diseños rebasan el alcance de este libro, pero en Montgomery (2008a) se encuentra su presentación.

Muchos de los conceptos que se analizaron en este capítulo se utilizarán en el capítulo 15, por ejemplo, la importancia de la aleatorización y el papel que desempeña la interacción en la interpretación de los resultados. Sin embargo, en el capítulo 15 se cubren 2 áreas que representan una expansión de los principios que se estudiaron en este capítulo y en el capítulo 13. En el capítulo 15 la solución de problemas con el uso de experimentos factoriales se realiza por medio del análisis de regresión, ya que se supone que la mayoría de los factores son cuantitativos y que se miden en un continuo, como la temperatura y el tiempo. Se derivan ecuaciones de predicción a partir de los datos del experimento diseñado y se utilizan para la mejora de procesos o incluso para su optimización. Además, se estudia el tema de los factoriales fraccionarios, en los que sólo una parte o fracción de todo el experimento factorial se aplica debido al costo excesivo que implica la realización de todo el experimento.

15. EXPERIMENTOS 2 FACTORIALES Y FRACCIONES

15.2.Introducción

Ya se han expuesto ciertos conceptos del diseño experimental. El plan de muestreo para la prueba t simple sobre la media de una población normal y el análisis de varianza implican la asignación aleatoria de los tratamientos preseleccionados a las unidades experimentales. El diseño de bloques aleatorizados, en el que los tratamientos se asignan a las unidades dentro de bloques relativamente homogéneos implica una aleatorización restringida.

En este capítulo se presta atención especial a los diseños experimentales en los que el plan experimental requiere estudiar el efecto sobre una respuesta de k factores, cada uno en dos niveles. A éstos se les conoce como **experimentos factoriales 2^k**. Es frecuente que los niveles se denoten por "alto" y "bajo", aunque esa notación podría ser arbitraria en el caso de variables cualitativas. El diseño factorial completo requiere que cada nivel de cada factor ocurra con cada nivel de cada uno de los demás factores, lo que da un total de **2^k combinaciones de tratamientos**.

15.3.Análisis de varianza

Considere inicialmente un factorial 2^2 con factores A y B, y *n observaciones experimentales por combinación de factores*. Es útil emplear los símbolos (1), a, b y ab para denotar los puntos del diseño, donde la presencia de una letra minúscula implica que el factor (A o B) está en el *nivel alto*. Así, la ausencia de la minúscula implica que el factor está en el *nivel bajo*. Por lo que ab es el punto de diseño $(+, +)$, a es $(+, -)$, b es $(-, +)$ y (1) es $(-, -)$. Asimismo existen situaciones en las que la notación también se aplica

para los datos de respuesta en el punto de diseño en cuestión. Como introducción al cálculo de **efectos** importantes que ayuden a determinar la influencia de los factores y **sumas de cuadrados** que están incorporados en los cálculos del análisis de varianza se presenta la tabla 15.1.

Tabla 15.1: Un experimento factorial 2^2

		A		**Media**
B		b	ab	$\frac{b+ab}{2n}$
		(1)	a	$\frac{(1)+a}{2n}$
Media		$\frac{(1)+b}{2n}$	$\frac{a+ab}{2n}$	

En esta tabla, (1), a, b y ab representan totales de los n valores de la respuesta en los puntos de diseño individuales. La simplicidad del factorial 2^2 reside en el hecho de que, aparte del error experimental, el analista obtiene la información importante en componentes con un solo grado de libertad, uno para cada uno de los dos efectos principales A y B, y un grado de libertad para la interacción AB. La información que se recupera sobre todos estos aspectos adopta la forma de tres **contrastes**. Se definirán los siguientes contrastes entre los totales de los tratamientos:

$$\begin{aligned} \text{contraste } A &= ab + a - b - (1), \\ \text{contraste } B &= ab - a + b - (1), \\ \text{contraste } AB &= ab - a - b + (1). \end{aligned}$$

Los tres **efectos** del experimento implican estos contrastes y apelan al sentido común y a la intuición. Los dos efectos principales calculados tienen la forma

$$\text{efecto} = \bar{y}_H - \bar{y}_L,$$

donde $\bar{y}_H$ y $\bar{y}_L$ son las respuestas promedio en el nivel alto o "+" y en el nivel bajo o "−", respectivamente. Como resultado,

Cálculo de los efectos principales

$$A = \frac{ab + a - b - (1)}{2n} = \frac{\text{contraste } A}{2n}$$

y

$$B = \frac{ab - a + b - (1)}{2n} = \frac{\text{contraste } B}{2n}.$$

La cantidad A es considerada la *diferencia entre la respuesta media en los niveles alto y bajo del factor A*. De hecho, A se denomina **efecto principal** del factor A. En forma similar, B es el efecto principal del factor B. Al inspeccionar la diferencia entre $ab - b$ y $a - (1)$ o entre $ab - a$ y $b - (1)$ en la tabla 15.1, se observa una aparente interacción en los datos. Si, por ejemplo,

$$ab - a \approx b - (1) \qquad \text{o bien} \qquad ab - a - b + (1) \approx 0,$$

Los números en las celdas de las tablas 15.2 y 15.3 ilustran con claridad la manera en que los contrastes y el cálculo resultante de los dos efectos principales y de las conclusiones resultantes pueden estar muy influidos por la presencia de interacción. En la tabla 15.2 el efecto de A es -30 tanto en el nivel bajo como en el nivel alto del factor B, y el efecto de B es 20 en los niveles bajo y alto del factor A. Esta "consistencia del efecto" (no hay interacción) puede ser información muy importante para el analista. Los efectos principales son

$$A = \frac{70+50}{2} - \frac{100+80}{2} = 60 - 90 = -30,$$
$$B = \frac{100+70}{2} - \frac{80+50}{2} = 85 - 65 = 20,$$

mientras que el efecto de la interacción es

$$AB = \frac{100+50}{2} - \frac{80+70}{2} = 75 - 75 = 0.$$

Por otro lado, en la tabla 15.3 el efecto A es nuevamente -30 al nivel bajo de B, pero $+30$ al nivel alto de B. Esta "inconsistencia del efecto" (interacción) también está presente para B en todos los niveles de A. En estos casos los efectos principales podrían carecer de significado y, de hecho, prestarse mucho a la confusión. Por ejemplo, el efecto de A es

$$A = \frac{50+70}{2} - \frac{80+40}{2} = 0,$$

ya que hay un "enmascaramiento" completo del efecto conforme se promedia sobre los niveles de B. La fuerte interacción se ilustra con el efecto calculado

$$AB = \frac{70+80}{2} - \frac{50+40}{2} = 30.$$

Aquí es conveniente ilustrar los escenarios de las tablas 15.2 y 15.3 con las gráficas de interacción. Observe el paralelismo en la gráfica de la figura 15.2 y la interacción aparente en la figura 15.3. ∎

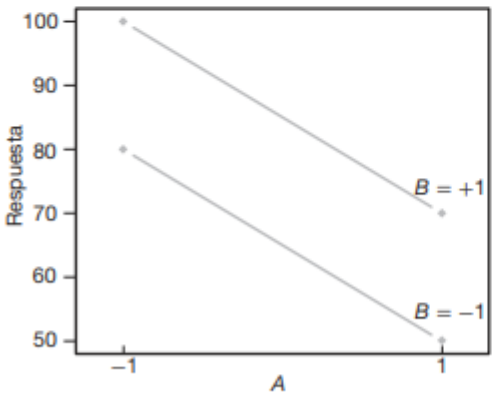


Figura 15.2: Gráfica de interacción para los datos de la tabla 15.2.

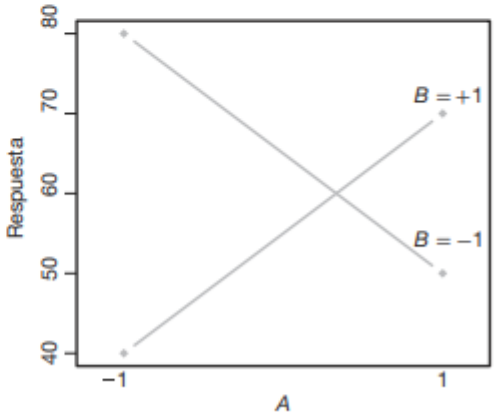


Figura 15.3: Gráfica de interacción para los datos de la tabla 15.3.

15.4.Experimento 2 K factorial sin replicas

El factorial completo 2^k con frecuencia requiere mucha experimentación, en particular cuando k es grande. Como resultado, a menudo no es posible replicar cada combinación de factores. Si en el modelo del experimento se incluyen todos los efectos, con todas las interacciones, no se permite ningún grado de libertad para el error. A menudo, cuando k es grande, el analista de datos *agrupará* las sumas de los cuadrados y los grados de libertad correspondientes para las interacciones de orden superior que se sabe, o se supone, son despreciables. Esto producirá pruebas F para los efectos principales e interacciones de orden inferior.

Graficación de diagnóstico con experimentos factoriales 2^k sin réplicas

Las gráficas de probabilidad normal constituyen una metodología muy útil para determinar la importancia relativa de los efectos en un experimento con factores de dos niveles razonablemente grandes cuando no hay réplica. Este tipo de gráfica de diagnóstico puede ser útil sobre todo cuando el analista de datos duda en agrupar interacciones de orden superior por temor de agrupar en el "error" algunos efectos verdaderamente reales y no sólo aleatorios. El lector debe recordar que todos los efectos que no son reales, es decir, que son *estimados de cero* independientes, siguen una distribución normal con media cercana a cero y varianza constante. Por ejemplo, en un experimento factorial 2^4 se debe recordar que todos los efectos, teniendo en cuenta que $n = 1$, son de la forma

$$AB = \frac{\text{contraste}}{8} = \bar{y}_H - \bar{y}_L,$$

donde $\bar{y}_H$ es el promedio de ocho corridas experimentales independientes en el nivel alto, o "+", y $\bar{y}_L$ es el promedio de ocho corridas independientes en el nivel bajo, o "−". Así, la varianza de cada contraste es $\text{Var}(\bar{y}_H - \bar{y}_L) = \sigma^2/4$. Para cualesquiera efectos reales $E(\bar{y}_H - \bar{y}_L) \neq 0$. Así, la gráfica de probabilidad normal debería revelar efectos "significativos" como aquellos que caen fuera de la línea recta que describe realizaciones de variables aleatorias normales independientes distribuidas de forma idéntica.

La gráfica de probabilidad puede adoptar una de muchas formas. Se recomienda al lector que consulte el capítulo 8, en el que se presentaron dichas gráficas por primera vez. Se puede usar la gráfica cuantil-cuantil, normal y empírica. También es posible utilizar el procedimiento de graficación que emplea el papel de probabilidad normal. Además, existen otros tipos de gráficas de probabilidad normal para el diagnóstico. En resumen, las gráficas de efectos para el diagnóstico son como sigue.

Gráficas de efectos de probabilidad para experimentos factoriales 2^4 sin réplica

1. Calcular los efectos como

$$\text{efecto} = \frac{\text{contraste}}{2^{k-1}}.$$

2. Construir una gráfica de probabilidad normal de todos los efectos.

3. Los efectos que caigan fuera de la línea recta deben considerarse reales.

A continuación se hacen más comentarios respecto de las gráficas de probabilidad normal de los efectos. En primer lugar, el analista podría sentirse frustrado si utiliza las gráficas con un experimento pequeño. Por otro lado, la graficación puede proporcionar resultados satisfactorios cuando hay *dispersión de efectos,* muchos efectos que no son verdaderamente reales. Esta dispersión será evidente en experimentos grandes, en los que es poco probable que las interacciones de orden superior sean reales.

Estudio de caso 15.1: **Moldeado por inyección.** Muchas empresas fabricantes de Estados Unidos y otros países utilizan partes moldeadas como componentes de un proceso. Un problema grande que enfrentan con frecuencia es el rebasamiento. A menudo, un molde troquelado de una parte se construye con un tamaño más grande que el nominal para permitir que se contraiga. En la siguiente situación experimental se produce un molde nuevo para el cual es importante encontrar las especificaciones adecuadas del proceso para minimizar la contracción. En el siguiente experimento los valores de la respuesta son desviaciones de los nominales, es decir, contracciones. Los factores y niveles son los siguientes:

	Niveles codificados	
	−1	**+1**
***A*. Velocidad de inyección (pies/seg)**	1.0	2.0
***B*. Temperatura de moldeado (°C)**	100	150
***C*. Presión de moldeado (psi)**	500	1000
***D*. Contrapresión (psi)**	75	120

El propósito del experimento fue determinar cuáles efectos (principales y de interacción) influyen en la contracción. El experimento se consideró un filtrado preliminar a partir del cual se determinaron los factores para un análisis más completo. Asimismo, se espera obtener información respecto a cómo los factores importantes repercuten en la contracción. En la tabla 15.6 se presentan los datos de un experimento factorial 2^4 sin réplica.

Tabla 15.6: Datos para el estudio de caso 15.1

Combinación de factores	**Respuesta (cm × 10^4)**	**Combinación de factores**	**Respuesta (cm × 10^4)**
(1)	72.68	*d*	73.52
a	71.74	*ad*	75.97
b	76.09	*bd*	74.28
ab	93.19	*abd*	92.87
c	71.25	*cd*	79.34
ac	70.59	*acd*	75.12
bc	70.92	*bcd*	79.67
abc	104.96	*abcd*	97.80

Inicialmente se calcularon los efectos y se plasmaron en una gráfica de probabilidad normal. Los efectos calculados son los siguientes:

$$\begin{aligned}
A &= 10.5613, & BD &= -2.2787, & B &= 12.4463,\\
C &= 2.4138, & D &= 2.1438, & AB &= 11.4038,\\
AC &= 1.2613, & AD &= -1.8238, & BC &= 1.8163,\\
CD &= 1.4088, & ABC &= 2.8588, & ABD &= -1.7813,\\
ACD &= -3.0438, & BCD &= -0.4788, & ABCD &= -1.3063.
\end{aligned}$$

En la figura 15.5 se observa la gráfica cuantil-cuantil normal, la cual parece implicar que los efectos *A*, *B* y *AB* son importantes. Los signos de los efectos importantes indican las conclusiones preliminares.

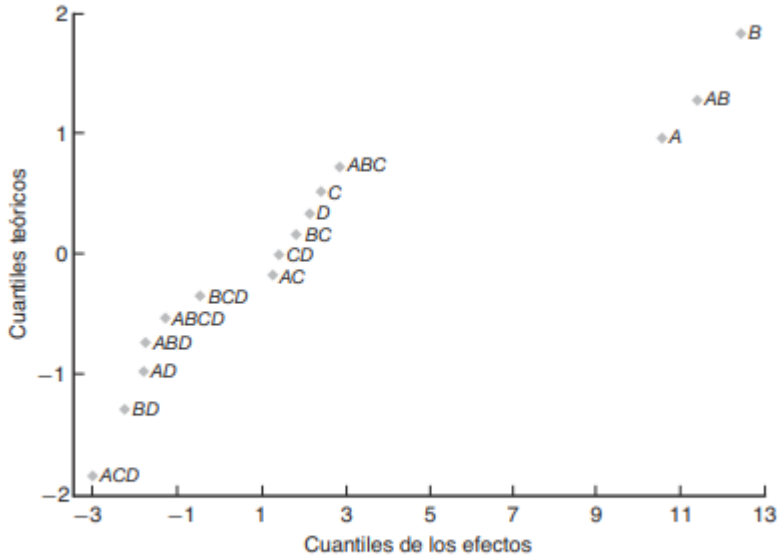


Figura 15.5: Gráfica cuantil-cuantil normal de los efectos para el estudio de caso del ejemplo 15.1.

1. Un incremento en la velocidad de inyección de 1.0 a 2.0 aumenta la contracción.
2. Un aumento en la temperatura de moldeado de 100°C a 150°C incrementa la contracción.
3. Hay una interacción entre la velocidad de inyección y la temperatura del moldeado; aunque ambos efectos principales son importantes es crucial entender el efecto de la interacción de los dos factores. ∎

Interpretación de la interacción de dos factores

Como se esperaría, una tabla de medias de dos factores facilita la interpretación de la interacción *AB*. Considere la situación de dos factores de la tabla 15.7.

Tabla 15.7: Ilustración de una interacción de dos factores

	***B* (temperatura)**	
***A* (velocidad)**	**100**	**150**
2	73.355	97.205
1	74.1975	75.240

Observe que la media muestral grande a velocidad y temperatura elevadas creó la interacción significativa. La **contracción se incrementa en forma no aditiva**. La temperatura del moldeado parece tener un efecto positivo a pesar del nivel de velocidad. Sin embargo, el efecto es el mayor a velocidad elevada. El efecto de la velocidad es muy ligero a temperaturas bajas, pero es claramente positivo a una temperatura elevada de moldeado. Para controlar la contracción a bajo nivel *debería evitarse el uso simultáneo de una alta velocidad de inyección y una temperatura de moldeado elevada*. Todos estos resultados se ilustran en forma gráfica en la figura 15.6.

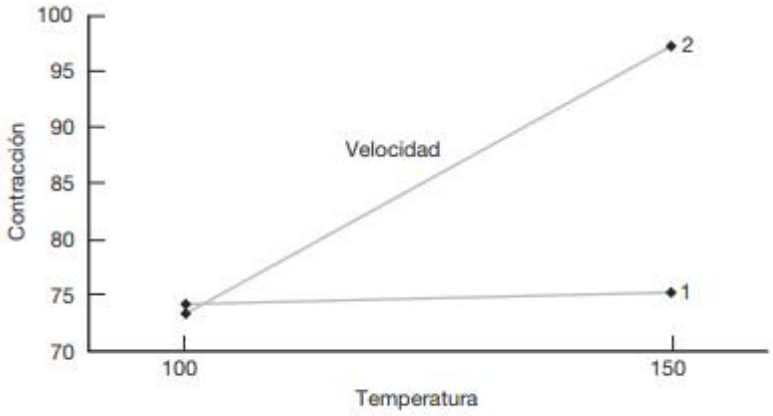


Figura 15.6: Gráfica de la interacción para el estudio de caso 15.1.

15.5.Experimentos factoriales de bloque sincompletos

Hasta ahora hemos limitado el análisis de los datos para un factorial 2^k al método del análisis de varianza. La única referencia a un análisis alternativo se hizo en el ejercicio 15.9 de la página 611. De hecho, este ejercicio introduce gran parte del material que da origen a la presente sección. Hay situaciones en las que el ajuste de un modelo es importante **y en la que es posible controlar** los factores que se estudian. Por ejemplo, un biólogo podría querer estudiar el crecimiento de cierto tipo de alga en el agua, en cuyo caso sería muy útil un modelo que relacionara las unidades de algas como una función de la *cantidad de cierto contaminante*, y, digamos, del *tiempo*. Así, el estudio involucra un experimento factorial en un ambiente de laboratorio en el que los factores son la concentración del contaminante y el tiempo. Como se verá más adelante en esta sección, es posible ajustar un modelo más preciso si los factores están controlados en un arreglo factorial, para el que con frecuencia es útil elegir un factorial 2^k. En muchos procesos biológicos y químicos los niveles de las variables regresoras pueden y deberían controlarse.

Hay que recordar que el modelo de regresión empleado en el capítulo 12 se puede escribir con notación de matriz de la siguiente manera

$$\mathbf{y} = \mathbf{X}\beta + \epsilon.$$

La matriz **X** se denomina **matriz del modelo**. Suponga, por ejemplo, que se utiliza un experimento factorial 2^3 con las variables

Temperatura:	150°C	200°C
Humedad:	15%	20%
Presión (psi):	1000	1500

Los niveles familiares $+1$ y -1 se generan a través del siguiente centrado y escalado a *unidades de diseño*:

$$x_1 = \frac{\text{temperatura} - 175}{25}, \quad x_2 = \frac{\text{humedad} - 17.5}{2.5}, \quad x_3 = \frac{\text{presión} - 1250}{250}.$$

Como resultado, la matriz **X** se vuelve

$$\mathbf{X} = \begin{array}{c} \begin{array}{cccc} & x_1 & x_2 & x_3 \end{array} \\ \begin{bmatrix} 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \end{array} \qquad \begin{array}{c} \text{Identificación del diseño} \\ (1) \\ a \\ b \\ c \\ ab \\ ac \\ bc \\ abc \end{array}$$

Ahora se observa que los contrastes ilustrados y analizados en la sección 15.2 están relacionados directamente con los coeficientes de regresión. Observe que todas las columnas de la matriz **X** en el ejemplo 2^3 son *ortogonales*. Como resultado, el cálculo de los coeficientes de regresión que se describió en la sección 12.3 se convierte en

$$b = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \left(\frac{1}{8}\mathbf{I}\right)\mathbf{X}'\mathbf{y}$$

$$= \frac{1}{8}\begin{bmatrix} a + ab + ac + abc + (1) + b + c + bc \\ a + ab + ac + abc - (1) - b - c - bc \\ b + ab + bc + abc - (1) - a - c - ac \\ c + ac + bc + abc - (1) - a - b - ab \end{bmatrix},$$

donde a, ab, etc., son medidas de la respuesta.

Ahora se observa que el concepto de *principales efectos calculados* que se enfatiza a lo largo de todo este capítulo con diseños factoriales 2^k, se relaciona con los coeficientes de un modelo de regresión ajustado cuando los factores son cuantitativos. De hecho, para un 2^k con, digamos, n corridas experimentales por punto del diseño, las relaciones entre los efectos y los coeficientes de regresión son como sigue:

$$\text{Efecto} = \frac{\text{contraste}}{2^{k-1}(n)}$$

$$\text{Coeficiente de regresión} = \frac{\text{contraste}}{2^k(n)} = \frac{\text{efecto}}{2}.$$

Esta relación debería tener sentido para el lector, ya que un coeficiente de regresión b_j es una tasa promedio del cambio en la respuesta *por cambio de unidad* en x_j. Por supuesto, cuando se va de -1 a $+1$ en x_j (de bajo a alto), la variable de diseño cambia en 2 unidades.

15.6.Diseño ortogonal

En situaciones experimentales en las que es apropiado ajustar modelos que son lineales en las variables de diseño y que posiblemente impliquen interacciones o términos de producto, el *diseño ortogonal* de dos niveles, o arreglo ortogonal, plantea algunas ventajas. Por diseño ortogonal nos referimos a uno en el que hay ortogonalidad entre las columnas de la matriz **X**. Considere la matriz **X** para el factorial 2^2 del ejemplo 15.2. Observe que las tres columnas son mutuamente ortogonales. La matriz **X** del factorial 2^3 también contiene columnas ortogonales. El factorial 2^3 con interacciones produciría una matriz **X** del tipo

$$\mathbf{X} = \begin{array}{c} \begin{array}{cccccccc} & x_1 & x_2 & x_3 & x_1x_2 & x_1x_3 & x_2x_3 & x_1x_2x_3 \end{array} \\ \begin{bmatrix} 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 \\ 1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \end{array}$$

La descripción de los grados de libertad es

Fuente	**g.l.**	
Regresión	3	
Falta de ajuste	4	$(x_1x_2, x_1x_3, x_2x_3, x_1x_2x_3)$
Error (puro)	8	
Total	15	

Los ocho grados de libertad para el error puro se obtienen a partir de las *corridas duplicadas* en cada punto del diseño. Los grados de libertad de la falta de ajuste podrían considerarse como la diferencia entre el número de puntos de diseño distintos y el número total de términos en el modelo; en este caso hay ocho puntos y cuatro términos en el modelo.

Error estándar de los coeficientes y pruebas *T*

En las secciones anteriores vimos cómo el diseñador de un experimento puede aprovechar el concepto de ortogonalidad para diseñar un experimento de regresión con coeficientes que obtienen una varianza mínima sobre la base del costo. Debemos ser capaces de utilizar el material sobre la regresión que se expuso en la sección 12.4 para calcular estimados de las varianzas de los coeficientes y, con ello, los errores estándar. También resulta de interés observar la relación entre el estadístico t de un coeficiente y el estadístico F descrito e ilustrado en capítulos anteriores.

En la sección 12.4 vimos que las varianzas y las covarianzas de los coeficientes aparecen en A^{-1}, o, en términos de la notación actual, la *matriz de varianza-covarianza* de coeficientes es

$$\sigma^2 A^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

En el caso del experimento factorial 2^k las columnas de **X** son mutuamente ortogonales,

lo que impone una estructura muy especial. En general, para 2^k se puede escribir

$$\mathbf{X} = [\mathbf{1} \quad \overset{x_1}{\pm \mathbf{1}} \quad \overset{x_2}{\pm \mathbf{1}} \quad \overset{\cdots}{\cdots} \quad \overset{x_k}{\pm \mathbf{1}} \quad \overset{x_1x_2}{\pm \mathbf{1}} \quad \overset{\cdots}{\cdots}],$$

donde cada columna contiene 2^k o $2^k n$ entradas, donde n es el número de réplicas de las corridas en cada punto del diseño. Así, la formación de $\mathbf{X}'\mathbf{X}$ lleva a

$$\mathbf{X}'\mathbf{X} = 2^k n \mathbf{I}_p,$$

donde **I** es la matriz de identidad de la dimensión p, el número de parámetros del modelo.

Una mirada más cercana a la propiedad de ortogonalidad del factorial 2^k

Ya vimos que para el caso del factorial 2^k toda la información que obtiene el analista sobre los efectos y las interacciones principales aparece en forma de contrastes. Estas "$2^k - 1$ piezas de información" conllevan un solo grado de libertad cada una y son independientes entre sí. En un análisis de varianza se manifiestan como *efectos*; mientras que si se construye un modelo de regresión, los efectos que resultan son coeficientes de regresión, aparte de un factor de 2. Con cada forma de análisis es posible hacer pruebas de significancia y la prueba t para un efecto dado es la misma en términos numéricos que para el coeficiente de regresión correspondiente. En el caso del ANOVA son importantes la selección de las variables y la interpretación científica de las interacciones; en tanto que en el caso de un análisis de regresión se usa un modelo para predecir la respuesta y/o determinar cuáles combinaciones de factores o niveles son las óptimas, por ejemplo, maximizar la producción o la eficiencia de la purificación, como en el estudio de caso 15.2.

Resulta que la propiedad de ortogonalidad es importante, ya sea que se trate de un ANOVA o de una regresión. La ortogonalidad entre las columnas de X, la matriz del modelo en, digamos, el ejemplo 15.3, ofrece condiciones especiales que tienen un impacto importante sobre los **efectos de la varianza** o los **coeficientes de regresión**. De hecho, ya es evidente que el diseño ortogonal da como resultado la igualdad de varianza para todos los efectos o coeficientes. Es así como, para propósitos de estimación o de prueba, la precisión es la misma para todos los coeficientes, los efectos principales o las interacciones. Además, si el modelo de regresión sólo contiene términos lineales, por lo cual sólo los efectos principales son de interés, las condiciones siguientes dan como resultado la minimización de las varianzas de todos los efectos, o, en forma correspondiente, de los coeficientes de regresión de primer orden.

Condiciones para varianzas mínimas de los coeficientes

Si el modelo de regresión contiene términos no mayores de primer orden, y si los rangos de las variables son dados por $x_j \in [-1, +1]$ para $j = 1, 2, \ldots, k$, entonces $\mathrm{Var}(b_j)/\sigma^2$, para $j = 1, 2, \ldots, k$, se minimiza si el diseño es ortogonal y todos los niveles x_i del diseño son ± 1 para $i = 1, 2, \ldots, k$.

Así, en términos de los coeficientes del modelo o los efectos principales, la ortogonalidad en el 2^k es una propiedad muy deseable.

Otro método para lograr una mejor comprensión del "balance" proporcionado por el factorial 2^3 consiste en observar la situación mediante una gráfica. En la figura 15.10 se aprecia cada uno de los contrastes ortogonales y, por lo tanto, mutuamente independientes. En las gráficas se comparan los planos de los cuadrados cuyos vértices contienen las respuestas etiquetadas con "+" con las que tienen el signo "−". Las que aparecen en el inciso *a* presentan contrastes para efectos principales y deberían ser evidentes para el lector. Las del inciso *b* presentan los planos determinados por los vértices "+" y "−" para los tres contrastes de interacción de dos factores. En el inciso *c* se aprecia la representación geométrica de los contrastes para la interacción de tres factores (ABC).

Corridas centrales con diseños factoriales 2^k

En la situación en que se aplica el diseño 2^k con variables **continuas** de diseño y se busca ajustar un modelo de regresión lineal, el uso de réplicas de corridas en el **diseño central** puede ser sumamente útil. De hecho, además de las ventajas que se analizarán a continuación, la mayoría de los científicos e ingenieros considerarían que las corridas

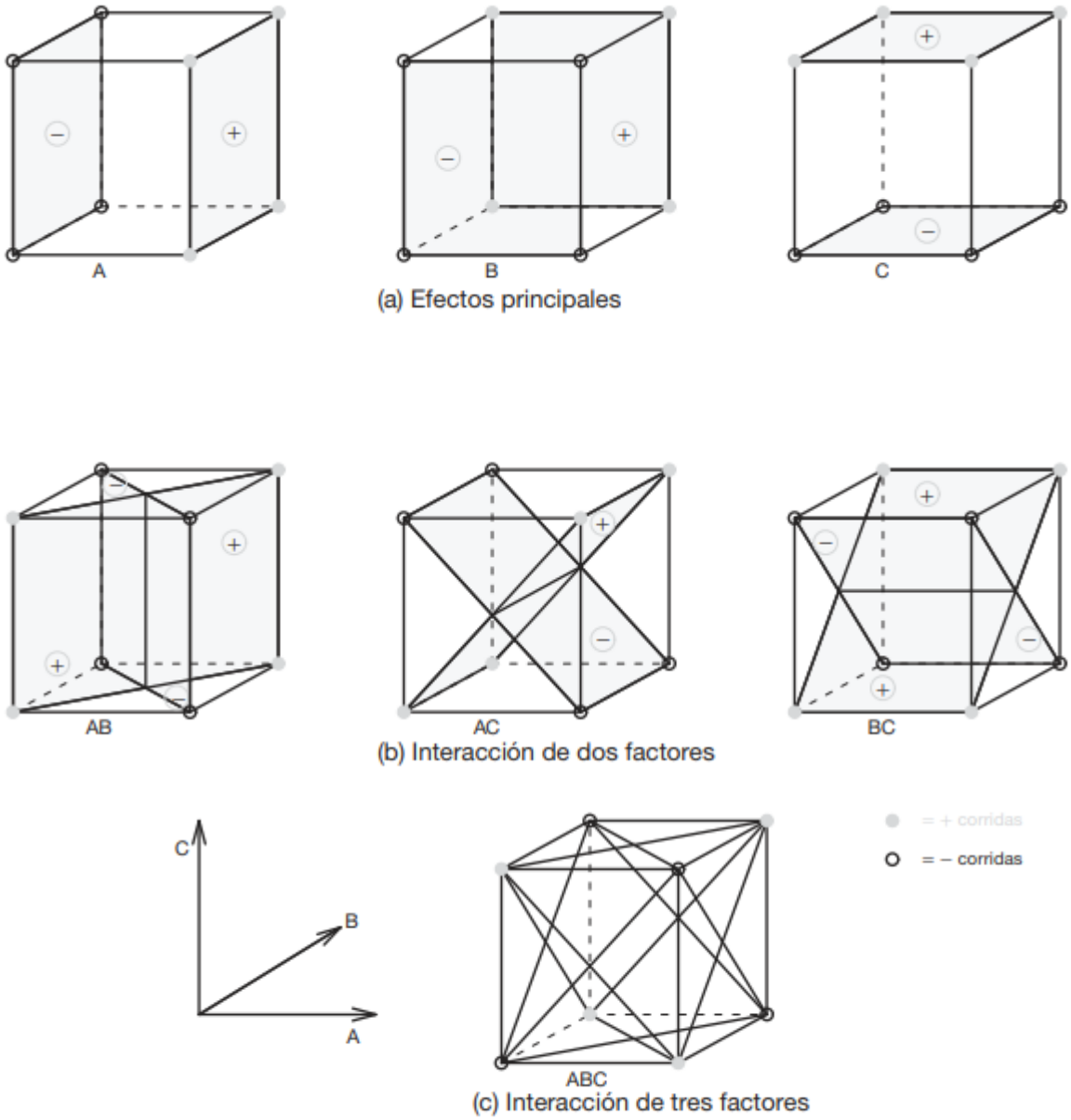


Figura 15.10: Presentación geométrica de los contrastes para el diseño factorial 2^3.

centrales, es decir, las corridas en $x_i = 0$ para $i = 1, 2,..., k$, no sólo son una práctica razonable sino que además son interesantes. En muchas áreas de aplicación del diseño 2^k el científico desea determinar si sería benéfico pasar a otra región de interés en los factores. En muchos casos el centro, es decir, el punto (0, 0,..., 0) en los factores codificados, con frecuencia representa las condiciones de operación actuales del proceso, o al menos aquellas condiciones que se consideran "óptimas para el momento". Por lo tanto, a menudo el científico requerirá datos sobre la respuesta central.

15.7.Experimentos factoriales fraccionados

El experimento factorial 2^k se puede volver muy demandante, en términos del número de unidades experimentales requeridas, cuando el valor de k es grande. Una de las ventajas reales de este plan experimental es que permite un grado de libertad para cada interacción. Sin embargo, en muchas situaciones experimentales se sabe que ciertas interacciones son despreciables, por lo que sería un desperdicio de esfuerzo experimental utilizar el experimento factorial completo. De hecho, el experimentador podría tener limitaciones económicas que le impidan hacer observaciones de todas las combinaciones 2^k de tratamientos. Cuando k es grande, a menudo se puede usar un **experimento factorial fraccionario** donde quizás sea posible llevar a cabo la mitad, un cuarto o incluso un octavo del plan factorial total.

Construcción de la fracción de $\frac{1}{2}$

La construcción del diseño de media réplica es idéntica a la asignación del experimento factorial 2^k en dos bloques. Se comienza por seleccionar un contraste de definición que se sacrificará por completo. Luego se construyen los dos bloques en concordancia y se elige cualquiera de ellos como plan experimental.

A menudo la fracción de $\frac{1}{2}$ de un factorial 2^k se conoce como diseño 2^{k-1}, el cual indica el número de puntos de diseño. El primer ejemplo de un diseño 2^{k-1} será uno de $\frac{1}{2}$ o uno de 2^3 o uno de 2^{3-1}. En otras palabras, el científico o el ingeniero no puede usar el complemento completo, es decir, todo el diseño 2^3 con 8 puntos de diseño, por lo que debe apelar a un diseño con sólo cuatro puntos de diseño. La pregunta es la siguiente: de los puntos de diseño (1), a, b, ab, ac, c, bc y abc, ¿cuáles son los cuatro puntos de diseño que producirán el diseño más útil? La respuesta, junto con los conceptos importantes relacionados, aparece en la tabla de signos + y − que muestra los contrastes para el diseño 2^3 completo. Considere la tabla 15.9.

Tabla 15.9: Contrastes para los siete efectos disponibles en el caso de un experimento factorial 2^3

	Combinación de tratamientos	Efectos							
		I	*A*	*B*	*C*	*AB*	*AC*	*BC*	*ABC*
2^{3-1}	a	+	+	−	−	−	−	+	+
	b	+	−	+	−	−	+	−	+
	c	+	−	−	+	+	−	−	+
	abc	+	+	+	+	+	+	+	+
2^{3-1}	ab	+	+	+	−	+	−	−	−
	ac	+	+	−	+	−	+	−	−
	bc	+	−	+	+	−	−	+	−
	(1)	+	−	−	−	+	+	+	−

Términos del modelo de regresión	3
Intersección	1
	4

Un análisis más detallado sugiere que los siete efectos no son ortogonales y que cada contraste está representado en otro efecto. De hecho, si se emplea el símbolo $\equiv$ para denotar **contrastes idénticos**, se tiene que

$$A \equiv BC; \qquad B \equiv AC; \qquad C \equiv AB.$$

Como resultado, dentro de un par no es posible estimar un efecto independiente de su "socio" alias. Los efectos

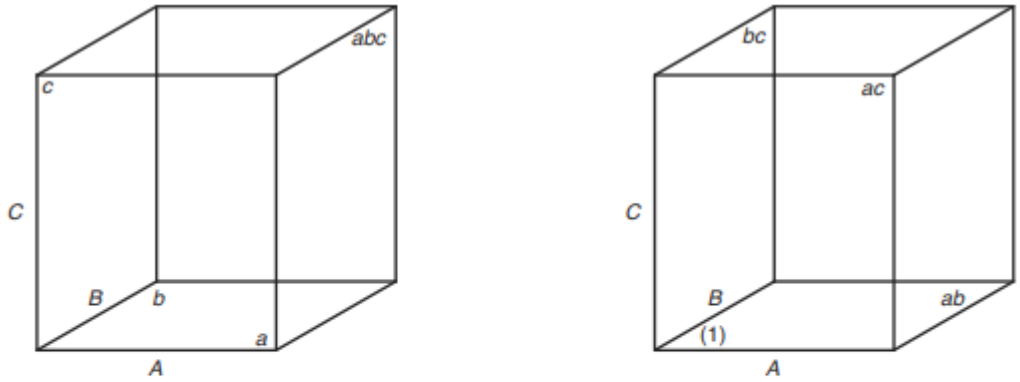
$$A = \frac{a + abc - b - c}{2} \qquad \text{y} \qquad BC = \frac{a + abc - b - c}{2}$$

producirán el mismo resultado numérico, de manera que contienen la misma información. De hecho, con frecuencia se dice que **comparten un grado de libertad**. En realidad, el efecto estimado verdaderamente estima la suma, es decir, $A + BC$. Se dice que A y BC son alias, al igual que B y AC, y que C y AB.

Para la fracción $ABC = -I$ se observa que los alias son los mismos que para la fracción $ABC = I$, además del signo. Así, se tiene

$$A \equiv -BC; \qquad B \equiv -AC; \qquad C \equiv -AB.$$

Las dos fracciones aparecen en las esquinas de los cubos de las figuras 15.15*a* y 15.15*b*.



(a) La fracción $ABC = I$ (a) La fracción $ABC = -I$

Figura 15.15: Las fracciones $\frac{1}{2}$ del factorial 2^3.

Cómo se determinan los alias en general

En general, para un diseño 2^{k-1}, cada efecto, además de aquel definido por el generador, tendrá un *solo socio alias*. El efecto definido por el generador no tendrá alias en otro

efecto, sino que su alias será la media, ya que el estimador de mínimos cuadrados será la media. Para determinar el alias de cada efecto, sólo se comienza con la relación definitoria, digamos $ABC = I$, para el diseño 2^{3-1}. Entonces, para obtener, digamos, el alias para el efecto A, se multiplica A por ambos lados de la ecuación $ABC = I$ y se reduce cualquier exponente por el módulo 2. Por ejemplo,

$$A \cdot ABC = A, \qquad \text{con lo que} \quad BC \equiv A.$$

En forma similar,

$$B \equiv B \cdot ABC \equiv AB^2C \equiv AC,$$

y, por supuesto,

$$C \equiv C \cdot ABC \equiv ABC^2 \equiv AB.$$

Ahora, para la segunda fracción, es decir, la definida por la relación $ABC = -I$,

$$A \equiv -BC; \qquad B \equiv -AC; \qquad C \equiv -AB.$$

Como resultado, el valor numérico del efecto A en realidad estima $A - BC$. De manera similar, el valor de B estima $B - AC$, y el valor de C estima $C - AB$.

15.8.Análisis de Experimentos factoriales fraccionados

La dificultad para realizar pruebas formales de significancia con datos de experimentos factoriales fraccionados radica en la determinación del término del error apropiado.

A menos que se disponga de datos de experimentos anteriores, el error debe provenir de una agrupación de contrastes que representan efectos que se presume son despreciables.

Las sumas de cuadrados para los efectos individuales se calculan usando en esencia los mismos procedimientos que se emplean para obtener el factorial completo. Es posible formar un contraste en las combinaciones de tratamientos construyendo la tabla de signos positivos y negativos. Por ejemplo, para media réplica de un experimento factorial 2^3 con ABC como contraste de definición, un conjunto posible de combinaciones de tratamientos, junto con el signo algebraico apropiado para cada contraste que se usa para calcular los efectos y las sumas de cuadrados de los distintos efectos, sería como el que se presenta en la tabla 15.13.

Tabla 15.13: Signos para los contrastes en media réplica de un experimento factorial 2^3

Combinación de tratamientos	**Efecto factorial**						
	A	***B***	***C***	***AB***	***AC***	***BC***	***ABC***
a	+	−	−	−	−	+	+
b	−	+	−	−	+	−	+
c	−	−	+	+	−	−	+
abc	+	+	+	+	+	+	+

Observe que en la tabla 15.13 los contrastes A y BC son idénticos, lo cual ilustra los alias. Asimismo, $B \equiv AC$ y $C \equiv AB$. En esta situación se tienen tres contrastes ortogonales que representan los 3 grados de libertad disponibles. Si se obtuvieran dos observaciones para cada una de las cuatro combinaciones de tratamientos, entonces tendríamos un estimado de la varianza del error con 4 grados de libertad. Si suponemos que los efectos de interacción son despreciables, podríamos probar la significancia de todos los efectos principales.

Un ejemplo del efecto y la suma de cuadrados correspondientes es

$$A = \frac{a - b - c + abc}{2n}, \qquad SCA = \frac{(a - b - c + abc)^2}{2^2 n}.$$

En general, la suma de cuadrados con un grado de libertad para cualquier efecto en una fracción 2^{-p} de un experimento factorial 2^k $(p < k)$ se obtiene elevando al cuadrado los contrastes en los totales de los tratamientos seleccionados y dividiendo entre $2^{k-p}n$, donde n es el número de réplicas de estas combinaciones de tratamientos.

15.9.Diseños de fracciones superiores y depurados

Algunas situaciones industriales requieren que el analista determine cuáles factores controlables, de entre un número grande de ellos, tienen un efecto sobre alguna respuesta importante. Los factores pueden ser cualitativos o variables de clase, variables de regresión o una mezcla de ambas. El procedimiento analítico puede requerir un análisis de varianza, una regresión o ambos. A menudo el modelo de regresión utilizado sólo incluye los efectos lineales principales, aunque tal vez sea posible estimar algunas interacciones. La situación exige la selección de variables y los diseños experimentales resultantes se denominan **diseños de filtrado**. Es evidente que los diseños ortogonales de dos niveles saturados o casi saturados son candidatos viables.

Resolución del diseño

A menudo los diseños ortogonales de dos niveles se clasifican según su **resolución**, la cual es determinada por la siguiente definición.

Definición 15.1: La **resolución** de un diseño ortogonal de dos niveles es la longitud de la interacción más pequeña (menos compleja) de entre el conjunto de contrastes de definición.

Si el diseño se construye como un factorial completo o fraccionado, ya sea un diseño 2^k, o bien, 2^{k-p}, $p = 1, 2, \ldots, k-1$, el concepto de resolución del diseño es un auxiliar para determinar el efecto de los alias. Por ejemplo, un diseño de resolución II sería de poca utilidad, ya que habría al menos un caso de alias de un efecto principal con otro. Un diseño de resolución III tendría todos sus efectos principales (lineales) ortogonales entre sí. No obstante, habrá algunos alias entre los efectos lineales y las interacciones de dos factores. Entonces, es evidente que si el analista está interesado en estudiar los efectos principales (lineales en el caso de la regresión) y no hay interacciones de dos factores, entonces se requiere un diseño cuya resolución sea de al menos III.

15.10. Construcción de diseños de resolución III y IV con 8 16 y 32 puntos de diseño

Es posible construir diseños útiles con resoluciones III y IV para 2 a 7 variables con 8 puntos de diseño. Empezamos con un factorial 2^3 que haya sido saturado simbólicamente con interacciones.

$$\begin{array}{ccccccc} x_1 & x_2 & x_3 & x_1x_2 & x_1x_3 & x_2x_3 & x_1x_2x_3 \end{array}$$
$$\begin{bmatrix} -1 & -1 & -1 & 1 & 1 & 1 & -1 \\ 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & 1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & 1 & -1 & -1 \\ -1 & 1 & 1 & -1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Es evidente que, con sólo reemplazar las columnas de interacción por nuevos efectos principales para las siete variables, se puede construir un diseño de resolución III. Por ejemplo, podríamos definir

$$\begin{aligned} x_4 &= x_1x_2 && \text{(contraste de definición } ABD) \\ x_5 &= x_1x_3 && \text{(contraste de definición } ACE) \\ x_6 &= x_2x_3 && \text{(contraste de definición } BCF) \\ x_7 &= x_1x_2x_3 && \text{(contraste de definición } ABCG) \end{aligned}$$

y obtendríamos una fracción 2^{-4} de un factorial 2^7. Las expresiones anteriores identifican los contrastes de definición elegidos. Resultan once contrastes de definición adicionales y todos contienen al menos tres letras. Así, el diseño es de resolución III. Es evidente que si se comienza con un *subconjunto* de columnas aumentadas y se concluye con un diseño

Tabla 15.16: Algunos diseños 2^{k-p} de resoluciones III, IV, V, VI y VII

Número de factores	Diseño	Número de puntos	Generadores
3	2_{III}^{3-1}	4	$C = \pm AB$
4	2_{IV}^{4-1}	8	$D = \pm ABC$
5	2_{III}^{5-2}	8	$D = \pm AB$; $E = \pm AC$
6	2_{VI}^{6-1}	32	$F = \pm ABCDE$
	2_{IV}^{6-2}	16	$E = \pm ABC$; $F = \pm BCD$
	2_{III}^{6-3}	8	$D = \pm AB$; $F = \pm BC$; $E = \pm AC$
7	2_{VII}^{7-1}	64	$G = \pm ABCDEF$
	2_{IV}^{7-2}	32	$E = \pm ABC$; $G = \pm ABDE$
	2_{IV}^{7-3}	16	$E = \pm ABC$; $F = \pm BCD$; $G = \pm ACD$
	2_{III}^{7-4}	8	$D = \pm AB$; $E = \pm AC$; $F = \pm BC$; $G = \pm ABC$
8	2_{V}^{8-2}	64	$G = \pm ABCD$; $H = \pm ABEF$
	2_{IV}^{8-3}	32	$F = \pm ABC$; $G = \pm ABD$; $H = \pm BCDE$
	2_{IV}^{8-4}	16	$E = \pm BCD$; $F = \pm ACD$; $G = \pm ABC$; $H = \pm ABD$

que incluye menos de 7 variables de diseño, el resultado es un diseño de resolución III en menos de siete variables.

Es posible construir un conjunto similar de diseños posibles para 16 puntos de diseño, comenzando con un diseño 2^4 saturado con interacciones. Las definiciones de las variables que corresponden a estas interacciones producen diseños de resolución III por medio de 15 variables. De manera similar, se pueden construir diseños que contengan 32 corridas, comenzando con un diseño 2^5.

La tabla 15.16 proporciona lineamientos para construir diseños de 8, 16, 32 y 64 puntos, con resolución III, IV e incluso V. La tabla proporciona el número de factores, el número de corridas y los generadores que se utilizan para producir los diseños 2^{k-p}. El generador dado se emplea para **aumentar el factorial completo** que contiene $k - p$ factores.

15.11. Otros diseños de resolución III de dos niveles diseños de Plackett-Burman

Una familia de diseños desarrollada por Plackett y Burman (1946, véase la bibliografía) llena el vacío del tamaño de la muestra que existe con los factoriales fraccionados. Éstos son útiles con muestras de tamaño 2^r, es decir, incluyen muestras de tamaños 4, 8, 16, 32, 64,... Los diseños de Plackett -Burman incluyen $4r$ puntos de diseño, por lo que se dispone de diseños de tamaño 12, 20, 24, 28, etcétera. Estos diseños de Plackett-Burman de dos niveles son diseños de resolución III y son muy fáciles de construir. Se proporcionan "renglones básicos" para cada tamaño de muestra. Estos renglones de signos + y − son $n - 1$ en número. Para construir las columnas de la matriz de diseño se comienza con el renglón básico y se hace una permutación cíclica sobre las columnas, hasta que se forman k columnas (el número deseado de variables). Después se llena el último renglón con signos negativos. El resultado será un diseño de resolución III con k variables ($k = 1, 2, \ldots, N$). Los renglones básicos son los siguientes:

$N = 12$	+ + − + + + − − − + −
$N = 16$	+ + + + − + − + + − − + − − −
$N = 20$	+ + − − + + + + − + − + − − − − + + −
$N = 24$	+ + + + + − + − + + − − + + − − + − + − − − −

Ejemplo 15.7 Construya un diseño depurado de dos niveles con 6 variables que contengan 12 puntos de diseño.

***Solución*:** Comience con el renglón básico en la columna inicial. La segunda columna se forma llevando la entrada inferior de la primera columna a la parte superior de la segunda, y repitiendo la primera. La tercera columna se forma del mismo modo, utilizando las entradas de la segunda columna. Cuando haya un número suficiente de columnas **sencillamente se llena el último renglón con signos negativos**. El diseño resultante es como sigue:

$$\begin{array}{c} \begin{array}{cccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{array} \\ \begin{bmatrix} + & - & + & - & - & - \\ + & + & - & + & - & - \\ - & + & + & - & + & - \\ + & - & + & + & - & + \\ + & + & - & + & + & - \\ + & + & + & - & + & + \\ - & + & + & + & - & + \\ - & - & + & + & + & - \\ - & - & - & + & + & + \\ + & - & - & - & + & + \\ - & + & - & - & - & + \\ - & - & - & - & - & - \end{bmatrix} \end{array}$$

Los diseños de Plackett-Burman son populares en la industria para situaciones de filtrado. Como se trata de diseños de resolución III. todos los efectos lineales son ortogonales. Para cualquier tamaño de muestra el usuario dispone de un diseño para $k = 2, 3, \ldots, N - 1$ variables.

La estructura de alias para el diseño de Plackett-Burman es muy complicada, por lo que el usuario no puede construir el diseño con un control completo de la estructura de alias, como en el caso de los diseños 2^k o 2^{k-p}. Sin embargo, en el caso de modelos de regresión el diseño de Plackett-Burman acepta interacciones (aunque no serán ortogonales) cuando se dispone de suficientes grados de libertad. ∎

15.12. Diseño robusto de parámetros de Taguchi

En el estudio de caso 15.2 se ajustó un modelo de regresión a un conjunto de datos con la meta específica de encontrar condiciones en esas variables de diseño que optimizaran (maximizaran) la eficiencia de purificación del carbón. El modelo incluía tres efectos principales lineales, tres términos de interacción de dos factores y un término de interacción de tres factores. La respuesta del modelo era la eficiencia de la purificación, y las condiciones óptimas de x_1, x_2 y x_3 se obtuvieron utilizando los signos y la magnitud de los coeficientes del modelo. En este ejemplo se utilizó un diseño de dos niveles para mejorar el proceso o para optimizarlo. En muchas áreas de la ciencia y de la ingeniería la aplicación se extiende para incluir modelos y diseños más complicados a los que, en conjunto, se les denomina **metodología de superficie de respuesta (MSR)**. Esta metodología abarca tanto métodos gráficos como analíticos. El término *superficie de respuesta* se deriva de la apariencia de la superficie multidimensional de la respuesta estimada constante de un modelo de segundo orden, es decir, un modelo con términos de primer y segundo orden. A continuación se presenta un ejemplo.

El modelo de superficie de respuesta de segundo orden

En muchos ejemplos industriales de optimización de procesos se utiliza un *modelo de superficie de respuesta de segundo orden*. Para el caso de, digamos $k = 2$ variables de proceso o variables de diseño, y una sola respuesta y, el modelo es dado por

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon.$$

Aquí se tienen $k = 2$ términos de primer orden, dos términos puros de segundo orden o cuadráticos y un término de interacción dado por $\beta_{12} x_1 x_2$. Los términos x_1 y x_2 se codifican en la forma conocida de ± 1. El término ϵ denota al acostumbrado error del modelo. En general, para k variables de diseño el modelo contendrá $1 + k + k + \binom{k}{2}$ términos del modelo y, por lo tanto, el diseño experimental debe contener al menos un número similar de puntos de diseño. Además, los términos cuadráticos requieren que las variables de diseño estén fijas en el diseño con al menos tres niveles. Al diseño resultante se le denomina *diseño de segundo orden*. A continuación se presenta un ejemplo.

El siguiente **diseño central compuesto (DCC)** y el ejemplo fueron tomados de Myers, Montgomery y Anderson-Cook (2009). Quizás la clase más popular de diseños de segundo orden sea la clase de los diseños centrales compuestos. El ejemplo que se presenta en la tabla 15.17 se refiere a un proceso químico en el que la temperatura de reacción, ξ_1, y la concentración del reactante, ξ_2, se muestran en sus niveles naturales y también de forma codificada. Cada factor tiene cinco niveles. Además, se incluye el orden en que se realizaron las observaciones de x_1 y x_2. La columna de la derecha proporciona los valores de la respuesta y, el porcentaje de conversión del proceso. Los primeros cuatro puntos de diseño representan los conocidos puntos factoriales en los niveles ± 1. Los siguientes cuatro puntos se conocen como puntos axiales, los cuales van seguidos por las corridas centrales que se explicaron y ejemplificaron antes en este capítulo. De esta manera, los cinco niveles de cada uno de los dos factores son -1, $+1$, -1.414, $+1.414$ y 0. En la figura 15.16 se presenta una imagen clara de la geometría del diseño central compuesto para este ejemplo de $k = 2$. En esta figura se ilustra la fuente del término **puntos axiales**. Estos cuatro puntos se localizan sobre los ejes factoriales, a una distancia axial de $\alpha = \sqrt{2} = 1.414$ a partir del centro del diseño. De hecho, para este DCC en particular, los puntos del perímetro, axiales y factoriales, se encuentran todos a la distancia $\sqrt{2}$ del centro del diseño, y como resultado tenemos ocho puntos equidistantes sobre un círculo más cuatro réplicas en el centro del diseño.

16. ESTADÍSTICA NO PARAMÉTRICA

16.1.Pruebas no paramétricas

La mayoría de los procedimientos de prueba de hipótesis que se presentaron en los capítulos anteriores se basan en la suposición de que las muestras aleatorias se seleccionan de poblaciones normales. Por fortuna la mayor parte de estas pruebas aún son confiables cuando existen ligeras desviaciones de la normalidad, en particular cuando el tamaño de la muestra es grande. Tradicionalmente, a tales procedimientos de prueba se les denomina **métodos paramétricos**. En este capítulo consideramos varios procedimientos de prueba alternativos, llamados **métodos no paramétricos** o **de distribución libre**, que a menudo no suponen conocimiento de ninguna clase acerca de las distribuciones de las poblaciones subyacentes, excepto, quizá, que éstas son continuas.

Los analistas de datos están usando procedimientos no paramétricos o de distribución libre cada vez con mayor frecuencia. En la ciencia y la ingeniería hay muchas aplicaciones en las que los datos no se reportan como valores de un continuo, sino, más bien, como una **escala ordinal** en la que es natural asignar rangos a los datos. De hecho, en este capítulo el lector notará muy pronto que los métodos de distribución libre aquí descritos implican un *análisis de rangos*. La mayoría de los analistas consideran que los cálculos involucrados en los métodos no paramétricos son muy atractivos e intuitivos.

Para revisar un ejemplo donde se aplica una prueba no paramétrica considere la situación en que dos jueces deben clasificar cinco marcas de cerveza de alta calidad asignando la categoría 1 a la marca que se considera que tiene la mejor calidad general, la categoría 2 a la segunda mejor, y así sucesivamente. Luego se puede utilizar una prueba no paramétrica para determinar si existe algún acuerdo entre los dos jueces.

También debemos señalar que las pruebas no paramétricas tienen asociadas varias desventajas. La primera es que no utilizan toda la información que proporciona la muestra, por lo tanto, cuando se pueden aplicar ambos métodos, estas últimas muestran ser menos eficientes que el procedimiento paramétrico correspondiente. En consecuencia, para lograr la misma potencia que la prueba paramétrica correspondiente, una prueba no paramétrica requerirá un tamaño muestral mayor que el que requeriría la primera.

Como antes indicamos, ligeras desviaciones de la normalidad dan como resultado desviaciones menores del ideal para las pruebas paramétricas estándar. Esto es particularmente cierto para la prueba t y la prueba F. En el caso de la prueba t y la prueba F, el

16.2. Prueba de signo

El lector debería recordar que los procedimientos que se estudiaron en la sección 10.4 para probar la hipótesis nula de que $\mu = \mu_0$ son válidos sólo si la población es aproximadamente normal o si la muestra es grande. Sin embargo, si $n < 30$ y la población decididamente no es normal, debemos recurrir a una prueba no paramétrica.

La prueba de signo se utiliza para probar hipótesis sobre una *mediana* de la población. En el caso de muchos de los procedimientos no paramétricos, la media es reemplazada por la mediana como el **parámetro de ubicación** pertinente a probar. Recuerde que la mediana muestral se definió en la sección 1.3. El equivalente de la población, que se denota con $\tilde{\mu}$, tiene una definición análoga. Dada una variable aleatoria X, $\tilde{\mu}$ se define de modo que $P(X > \tilde{\mu}) \leq 0.5$ y $P(X < \tilde{\mu}) \leq 0.5$. En el caso continuo,

$$P(X > \tilde{\mu}) = P(X < \tilde{\mu}) = 0.5.$$

Por supuesto, si la distribución es simétrica, la media y la mediana de la población son iguales. Al probar la hipótesis nula H_0 de que $\tilde{\mu} = \tilde{\mu}_0$ en comparación con la hipótesis alternativa adecuada, con base en una muestra aleatoria de tamaño n, reemplazamos cada valor de la muestra que exceda a $\tilde{\mu}_0$ con un signo *más*, y cada valor de la muestra menor que $\tilde{\mu}_0$ con un signo *menos*. Si la hipótesis nula es verdadera y la población es simétrica, la suma de los signos más debería ser casi igual a la suma de los signos menos. Cuando un signo aparece con más frecuencia de lo que debería, con base sólo en el azar, rechazamos la hipótesis de que la mediana de la población $\tilde{\mu}$ es igual a $\tilde{\mu}_0$.

En teoría, la prueba de signo sólo se puede aplicar en situaciones en las que $\tilde{\mu}_0$ no puede ser igual al valor de cualquiera de las observaciones. Aunque la probabilidad de obtener una observación muestral exactamente igual a $\tilde{\mu}_0$ cuando la población es continua es de cero, en la práctica un valor de la muestra igual a $\tilde{\mu}_0$ ocurre con frecuencia debido a una falta de precisión en el registro de los datos. Cuando se observan valores de la muestra iguales a $\tilde{\mu}_0$, se excluyen del análisis, lo cual da como resultado que se reduzca el tamaño de la muestra.

El estadístico de prueba adecuado para la prueba de signo es la variable aleatoria binomial X, que representa el número de signos más en la muestra aleatoria. Si la hipótesis nula de que $\tilde{\mu} = \tilde{\mu}_0$ es verdadera, la probabilidad de que un valor muestral dé como resultado un signo más o uno menos es igual a $1/2$. Por lo tanto, para probar la hipótesis nula de que $\tilde{\mu} = \tilde{\mu}_0$, en realidad probamos la hipótesis nula de que el número de signos

más es un valor de una variable aleatoria que tiene una distribución binomial con el parámetro $p = 1/2$. Por lo tanto, los valores P para las alternativas unilateral y bilateral se pueden calcular usando esta distribución binomial. Por ejemplo, probando

$$H_0\colon\ \tilde{\mu} = \tilde{\mu}_0,$$
$$H_1\colon\ \tilde{\mu} < \tilde{\mu}_0,$$

se rechaza H_0 a favor de H_1 sólo si la proporción de signos más es lo suficientemente menor que $1/2$, es decir, cuando el valor x de la variable aleatoria es pequeño. Por lo tanto, si el valor P que se calcula

$$P = P\,(X \leq x \text{ cuando } p = 1/2)$$

es menor o igual que algún nivel de significancia α preseleccionado, se rechaza H_0 a favor de H_1. Por ejemplo, cuando $n = 15$ y $x = 3$, en la tabla A.1 encontramos que

$$P = P(X \leq 3 \text{ cuando } p = 1/2) = \sum_{x=0}^{3} b\left(x; 15, \frac{1}{2}\right) = 0.0176,$$

de manera que la hipótesis nula $\tilde{\mu} = \tilde{\mu}_0$ realmente se puede rechazar a un nivel de significancia de 0.05 pero no a un nivel de 0.01.

Para probar la hipótesis

$$H_0\colon\ \tilde{\mu} = \tilde{\mu}_0,$$
$$H_1\colon\ \tilde{\mu} > \tilde{\mu}_0,$$

se rechaza H_0 a favor de H_1 sólo si la proporción de signos más es suficientemente mayor que $1/2$, es decir, cuando x es grande. En consecuencia, si el valor P calculado

$$P = P\,(X \geq x \text{ cuando } p = 1/2)$$

es menor que α, se rechaza H_0 a favor de H_1. Finalmente, para probar la hipótesis

Siempre que $n > 10$, las probabilidades binomiales con $p = 1/2$ se pueden aproximar a partir de la curva normal, ya que $np = nq > 5$. Suponga, por ejemplo, que deseamos probar la hipótesis

$$H_0\colon\ \tilde{\mu} = \tilde{\mu}_0,$$
$$H_1\colon\ \tilde{\mu} < \tilde{\mu}_0,$$

a un nivel de significancia $\alpha = 0.05$ para una muestra aleatoria de tamaño $n = 20$ que produce $x = 6$ signos más. Si utilizamos la aproximación de la curva normal con

$$\tilde{\mu} = np = (20)(0.5) = 10$$

y

$$\sigma = \sqrt{npq} = \sqrt{(20)(0.5)(0.5)} = 2.236,$$

encontramos que

$$z = \frac{6.5 - 10}{2.236} = -1.57.$$

Por lo tanto,

$$P = P(X \leq 6) \approx P(Z < -1.57) = 0.0582,$$

que conduce a no rechazar la hipótesis nula.

16.3.Prueba de rango con signo

El lector debe notar que la prueba de signo sólo utiliza los signos más y menos de las diferencias entre las observaciones y $\tilde{\mu}_0$ en el caso de una muestra, o los signos más y menos de las diferencias entre los pares de observaciones en el caso de muestras en pares; no se toma en cuenta la magnitud de esas diferencias. Una prueba que utiliza dirección y magnitud, propuesta en 1945 por Frank Wilcoxon, ahora se conoce comúnmente como **prueba de rango con signo de Wilcoxon**.

El analista puede extraer más información de los datos de manera no paramétrica si es razonable aplicar una restricción adicional a la distribución de la que se toman los datos. La prueba de rango con signo de Wilcoxon se aplica en el caso de una **distribución continua simétrica**. En esta condición se prueba la hipótesis nula $\tilde{\mu} = \tilde{\mu}_0$. Primero restamos $\tilde{\mu}_0$ de cada valor muestral y descartamos todas las diferencias iguales a cero. Las diferencias restantes se ordenan sin importar el signo. Se asigna una categoría de 1a la diferencia absoluta más pequeña, es decir, sin signo, una categoría de 2 a la siguiente más pequeña, y así sucesivamente. Cuando el valor absoluto de dos o más diferencias es el mismo, se asigna a cada uno el promedio de los rangos que se asignarían si las diferencias fueran distinguibles. Por ejemplo, si la quinta y la sexta diferencias más pequeñas tienen el mismo valor absoluto, a cada una se le asignaría una categoría de 5.5. Si la hipótesis $\tilde{\mu} = \tilde{\mu}_0$ es verdadera, el total de los rangos que corresponden a las diferencias positivas debería ser casi igual al total de los rangos que corresponden a las diferencias negativas. Representemos estos totales con w_+ y w_-, respectivamente. Designamos el más pequeño de w_+ y w_- con w.

Al seleccionar muestras repetidas esperaríamos que w_+ y w_- y, por lo tanto, w variará. De esta manera, consideramos a w_+, w_- y w como valores de las correspondientes variables aleatorias W_+, W_- y W. La hipótesis nula $\tilde{\mu} = \tilde{\mu}_0$ se puede rechazar a favor de la hipótesis alternativa $\tilde{\mu} < \tilde{\mu}_0$ sólo si w_+ es pequeña y w_- es grande. De igual manera, la hipótesis alternativa $\tilde{\mu} > \tilde{\mu}_0$ se puede aceptar sólo si w_+ es grande y w_- es pequeña. Para una alternativa bilateral se puede rechazar H_0 a favor de H_1 si w_+ o w_- y, en consecuencia, w son suficientemente pequeñas. Por lo tanto, no importa cuál sea la hipótesis alternativa,

cuando el valor del estadístico adecuado W_+, W_- o W es suficientemente pequeño, se rechaza la hipótesis nula.

Dos muestras con observaciones en pares

Con el fin de probar la hipótesis nula de que se toman muestras de dos poblaciones simétricas continuas con $\tilde{\mu}_1 = \tilde{\mu}_2$ para el caso de muestras en pares, se ordenan las diferencias de las observaciones en pares sin importar el signo y se procede como en el caso de una sola muestra. Los diversos procedimientos de prueba para los casos de una sola muestra y de muestras en pares se resumen en la tabla 16.2.

Tabla 16.2: Prueba de rango con signo

H_0	H_1	Calcular
$\tilde{\mu} = \tilde{\mu}_0$	$\tilde{\mu} < \tilde{\mu}_0$	w_+
	$\tilde{\mu} > \tilde{\mu}_0$	w_-
	$\tilde{\mu} \neq \tilde{\mu}_0$	w
$\tilde{\mu}_1 = \tilde{\mu}_2$	$\tilde{\mu}_1 < \tilde{\mu}_2$	w_+
	$\tilde{\mu}_1 > \tilde{\mu}_2$	w_-
	$\tilde{\mu}_1 \neq \tilde{\mu}_2$	w

No es difícil mostrar que siempre que $n < 5$ y que el nivel de significancia no exceda a 0.05 para una prueba de una cola, o a 0.10 para una prueba de dos colas, todos los valores posibles de w_+, w_- o w conducirán a la aceptación de la hipótesis nula. Sin embargo, cuando $5 \leq n \leq 30$, la tabla A.16 muestra valores críticos aproximados de W_+ y W_- a niveles de significancia iguales a 0.01, 0.025 y 0.05 para una prueba de una cola y valores críticos de W a niveles de significancia iguales a 0.02, 0.05 y 0.10 para una prueba de dos colas. Se rechaza la hipótesis nula si el valor calculado w_+, w_- o w es **menor o igual que** el valor tabulado apropiado. Por ejemplo, cuando $n = 12$, la tabla A.16 indica que se requiere un valor de $w_+ \leq 17$ para que la hipótesis alternativa unilateral $\tilde{\mu} < \tilde{\mu}_0$ sea significativa al nivel 0.05.

16.4.Prueba de la suma de rangos

Como antes indicamos, el procedimiento no paramétrico por lo general es una alternativa adecuada para la prueba de la teoría normal cuando la suposición de normalidad no es válida. Cuando nos interesa probar la igualdad de las medias de dos distribuciones continuas que evidentemente no son normales, y las muestras son independientes, es decir, que no hay emparejamiento de observaciones, la **prueba de la suma de rangos de Wilcoxon** o la **prueba de dos muestras de Wilcoxon** es una alternativa apropiada a la prueba t de dos muestras que se describe en el capítulo 10.

Probaremos la hipótesis nula H_0 de que $\tilde{\mu}_1 = \tilde{\mu}_2$ en comparación con alguna hipótesis alternativa adecuada. Primero seleccionamos una muestra aleatoria de cada una de las poblaciones. Sea n_1 el número de observaciones en la muestra más pequeña y n_2 el número de observaciones en la muestra más grande. Cuando las muestras son de igual tamaño n_1 y n_2 se pueden asignar de manera aleatoria. Se ordenan las $n_1 + n_2$ observaciones de las muestras combinadas en orden ascendente y se sustituye un rango de 1, 2,..., $n_1 + n_2$ para cada observación. En el caso de empates (observaciones idénticas), se reemplazan las observaciones por la media de los rangos que tendrían las observaciones si fueran distinguibles. Por ejemplo, si la séptima y octava observaciones fueran idénticas, se asignaría un rango de 7.5 a cada una de las dos observaciones.

La suma de los rangos que corresponden a las n_1 observaciones en la muestra más pequeña se denota con w_1. De manera similar, el valor w_2 representa la suma de los n_2 rangos que corresponden a la muestra más grande. El total $w_1 + w_2$ depende sólo del número de observaciones en las dos muestras y de ninguna manera resulta afectado por los resultados del experimento. Por lo tanto, si $n_1 = 3$ y $n_2 = 4$, entonces $w_1 + w_2 = 1 + 2 + \cdots + 7 = 28$, sin importar los valores numéricos de las observaciones. En general,

$$w_1 + w_2 = \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2},$$

la suma aritmética de los enteros 1, 2,..., $n_1 + n_2$. Una vez que se determina w_1, es más fácil calcular w_2 mediante la fórmula

$$w_2 = \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2} - w_1.$$

que U_1 y U_2 tienen distribuciones muestrales simétricas y toman valores en el intervalo de 0 a n_1n_2, tales que $u_1 + u_2 = n_1n_2$.

De las fórmulas para u_1 y u_2 vemos que u_1 será pequeña cuando w_1 es pequeña, y u_2 será pequeña cuando w_2 sea pequeña. En consecuencia, la hipótesis nula se rechazará siempre que los estadísticos apropiados U_1, U_2 o U tomen un valor menor o igual que el valor crítico deseado dado en la tabla A.17. Los diversos procedimientos de prueba se resumen en la tabla 16.4.

Tabla 16.4: Prueba de la suma de rangos

H_0	H_1	Calcular
$\tilde{\mu}_1 = \tilde{\mu}_2$	$\tilde{\mu}_1 < \tilde{\mu}_2$	u_1
	$\tilde{\mu}_1 > \tilde{\mu}_2$	u_2
	$\tilde{\mu}_1 \neq \tilde{\mu}_2$	u

La tabla A.17 proporciona valores críticos de U_1 y U_2 para niveles de significancia iguales a 0.001, 0.01, 0.025 y 0.05 para una prueba de una cola, y valores críticos de U para niveles de significancia iguales a 0.002, 0.02, 0.05 y 0.10 para una prueba de dos colas. Si el valor observado de u_1, u_2 o u es **menor o igual que** el valor crítico tabulado, se rechaza la hipótesis nula al nivel de significancia que se indica en la tabla. Suponga, por ejemplo, que deseamos probar la hipótesis nula de que $\tilde{\mu}_1 = \tilde{\mu}_2$ en comparación con la hipótesis alternativa unilateral de que $\tilde{\mu}_1 < \tilde{\mu}_2$ a un nivel de significancia de 0.05 para muestras aleatorias de tamaños $n_1 = 3$ y $n_2 = 5$, que producen el valor $w_1 = 8$. Se sigue que

$$u_1 = 8 - \frac{(3)(4)}{2} = 2.$$

Nuestra prueba de una sola cola se basa en el estadístico U_1. Si se usa la tabla A.17, se rechaza la hipótesis nula de medias iguales cuando $u_1 \leq 1$. Como $u_1 = 2$ no cae en la región de rechazo, no se puede rechazar la hipótesis nula.

Teoría normal de aproximación para dos muestras

Cuando n_1 y n_2 exceden a 8, la distribución muestral de U_1 (o U_2) se aproxima a la distribución normal con media y varianza dadas por

$$\mu_{U_1} = \frac{n_1n_2}{2} \text{ y } \sigma^2_{U_1} = \frac{n_1n_2(n_1 + n_2 + 1)}{12}.$$

En consecuencia, cuando n_2 es mayor que 20, el valor máximo en la tabla A.17, y n_1 es al menos 9, se puede utilizar el estadístico

$$Z = \frac{U_1 - \mu_{U_1}}{\sigma_{U_1}}$$

para la prueba, con la región crítica que cae ya sea en alguna o en ambas colas de la distribución normal estándar, dependiendo de la forma de H_1.

El uso de la prueba de suma de rangos de Wilcoxon no se restringe a poblaciones no normales. Se puede utilizar en vez de la prueba t de dos muestras cuando las poblaciones son normales, aunque la potencia será menor. La prueba de suma de rangos de Wilcoxon siempre es superior a la prueba t para poblaciones definitivamente no normales.

16.5.Prueba de Kruskal Wallis

En los capítulos 13, 14 y 15 la técnica del análisis de varianza resalta como técnica analítica para probar la igualdad de $k \geq 2$ medias de la población. Sin embargo, el lector debería recordar que para que la prueba F sea teóricamente correcta se debe suponer normalidad. En esta sección investigamos una alternativa no paramétrica al análisis de varianza.

La **prueba de Kruskal-Wallis**, también llamada **prueba *H* de Kruskal-Wallis**, es una generalización de la prueba de la suma de rangos para el caso de $k > 2$ muestras. Se utiliza para probar la hipótesis nula H_0 de que k muestras independientes provienen de poblaciones idénticas. Presentada en 1952 por W. H. Kruskal y W. A. Wallis, la prueba constituye un procedimiento no paramétrico para probar la igualdad de las medias, en el análisis de varianza de un factor, cuando el experimentador desea evitar la suposición de que las muestras se seleccionaron de poblaciones normales.

Sea n_i $(i = 1, 2,..., k)$ el número de observaciones en la i-ésima muestra. Primero combinamos todas las k muestras y acomodamos las $n = n_1 + n_2 + \cdots + n_k$ observaciones en orden ascendente, y sustituimos el rango apropiado de $1, 2,..., n$ para cada observación. En el caso de empates (observaciones idénticas), seguimos el procedimiento acostumbrado de reemplazar las observaciones por la media de los rangos que tendrían las observaciones si fueran distinguibles. La suma de los rangos que corresponde a las n_i observaciones en la i-ésima muestra se denota mediante la variable aleatoria R_i. Consideremos ahora el estadístico

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1),$$

que se aproxima muy bien mediante una distribución chi cuadrada con $k - 1$ grados de libertad, cuando H_0 es verdadera, siempre y cuando cada muestra conste de al menos 5 observaciones. El hecho de que h, el supuesto valor de H, sea grande cuando las muestras independientes provienen de poblaciones que no son idénticas nos permite establecer el siguiente criterio de decisión para probar H_0:

Prueba de Kruskal-Wallis

Para probar la hipótesis nula H_0 de que k muestras independientes provienen de poblaciones idénticas se calcula

$$h = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{r_i^2}{n_i} - 3(n+1),$$

donde r_i es el valor supuesto de R_i para $i = 1, 2,..., k$. Si h cae en la región crítica $H > \chi^2_\alpha$ con $v = k - 1$ grados de libertad, se rechaza H_0 al nivel de significancia α; de otra manera no se rechaza H_0.

16.6.Pruebas de rachas

Al aplicar los diversos conceptos estadísticos que se presentan a lo largo de este libro siempre asumimos que los datos muestrales se reunieron mediante algún procedimiento aleatorio. Las **pruebas de rachas**, que se basan en el orden en el que se obtienen las observaciones muestrales, constituyen una técnica útil para probar la hipótesis nula H_0 de que las observaciones en realidad se obtuvieron al azar.

Para ilustrar las pruebas de rachas suponga que se encuesta a 12 personas para saber si utilizan cierto producto. Se cuestionaría seriamente la supuesta aleatoriedad de la muestra si las 12 personas fueran del mismo sexo. Designaremos a un hombre y a una mujer con los símbolos H y M, respectivamente, y registraremos los resultados de acuerdo con su género en el orden en que ocurren. Una secuencia común para el experimento sería

$$\underbrace{M\ M}\ \underbrace{F\ F\ F}\ \underbrace{M}\ \underbrace{F\ F}\ \underbrace{M\ M\ M\ M},$$

donde agrupamos las subsecuencias de símbolos idénticos. Tales agrupamientos se llaman **rachas**.

Definición 16.1: Una **racha** es una subsecuencia de uno o más símbolos idénticos que representan una propiedad común de los datos.

Sin importar si las mediciones de la muestra representan datos cualitativos o cuantitativos, la prueba de rachas divide los datos en dos categorías mutuamente excluyentes: hombre o mujer, defectuoso o no defectuoso, cara o cruz, arriba o abajo de la mediana, etcétera. En consecuencia, una secuencia siempre estará limitada a dos símbolos distintos. Sea n_1 el número de símbolos asociados con la categoría de menor ocurrencia, y n_2 el número de símbolos que pertenecen a la otra categoría. Entonces, el tamaño de la muestra $n = n_1 + n_2$.

Para los $n = 12$ símbolos en nuestra encuesta tenemos cinco rachas, donde la primera incluye dos H, la segunda tres M, y así sucesivamente. Si el número de rachas es mayor o menor que el que esperaríamos por el azar, se debe rechazar la hipótesis de que la muestra se extrajo al azar. Ciertamente, una muestra que tiene como resultado sólo dos corridas,

$$H\,H\,H\,H\,H\,H\,H\,M\,M\,M\,M\,M$$

o la inversa, es muy improbable que provenga de un proceso de selección aleatorio. Este resultado indicaría que las primeras siete personas entrevistadas son hombres, seguidos de cinco mujeres. Asimismo, si la muestra tiene como resultado el número máximo de 12 rachas, como en la secuencia alternada

$$H\,M\,H\,M\,H\,M\,H\,M\,H\,M\,H\,M,$$

de nuevo sospecharíamos del orden en que se seleccionaron los individuos para la encuesta.

La prueba de rachas para la aleatoriedad se basa en la variable aleatoria V, el número total de rachas que suceden en la secuencia completa del experimento. En la tabla A.18 se dan valores de $P(V \leq v^*$ cuando H_0 es verdadera) para $v^* = 2, 3, \ldots, 20$ rachas y valores de n_1 y n_2 menores o iguales que 10. Los valores P tanto para pruebas de una cola como de dos colas se pueden obtener usando estos valores tabulados.

En la encuesta anterior presentamos un total de 5 M y 7 H. De aquí, con $n_1 = 5$, $n_2 = 7$ y $v = 5$, en la tabla A.18 observamos que el valor P para una prueba de dos colas es

$$P = 2P(V \leq 5 \text{ cuando } H_0 \text{ es verdadera}) = 0.394 > 0.05.$$

Es decir, el valor $v = 5$ es razonable a un nivel de significancia de 0.05 cuando H_0 es verdadera y, por lo tanto, no tenemos suficiente evidencia para rechazar la hipótesis de aleatoriedad de nuestra muestra.

Cuando el número de rachas es grande, por ejemplo, cuando $v = 11$ y $n_1 = 5$ y $n_2 = 7$, entonces el valor P en una prueba de dos colas es

$$\begin{aligned} P &= 2P(V \geq 11 \text{ cuando } H_0 \text{ es verdadera}) = 2[1 - P(V \leq 10 \text{ cuando } H_0 \text{ es verdadera})] \\ &= 2(1 - 0.992) = 0.016 < 0.05, \end{aligned}$$

que nos lleva a rechazar la hipótesis de que los valores de la muestra ocurren al azar.

La prueba de rachas también sirve para detectar desviaciones en la aleatoriedad de una secuencia de mediciones cuantitativas a lo largo del tiempo, ocasionadas por tendencias o periodos. Al reemplazar cada medición en el orden en que se obtiene, con un símbolo *más* si caen por arriba de la mediana, o con un símbolo *menos* si caen por debajo de la mediana, y omitiendo todas las mediciones que son exactamente iguales a la mediana, se genera una secuencia de signos de más y menos que se somete a prueba para verificar su aleatoriedad, como se ilustra en el siguiente ejemplo.

16.7.Límites de tolerancia

En el capítulo 9 se analizaron los límites de tolerancia para una distribución normal de mediciones. En esta sección consideramos un método para construir intervalos de tolerancia que sean independientes de la forma de la distribución subyacente. Como se podría sospechar, para un grado de confianza razonable serán considerablemente más grandes que los que se construyen cuando se supone normalidad, y el tamaño de la muestra que se requiere es por lo general muy grande. Los límites de tolerancia no paramétricos se establecen en términos de las observaciones más grande y más pequeña en nuestra muestra.

Límites de tolerancia bilaterales

Para cualquier distribución de mediciones los límites de tolerancia bilaterales son indicados por las observaciones más grande y más pequeña en una muestra de tamaño n, donde n se determina de manera que se asegure, con $100(1-\gamma)\%$ de confianza, que **al menos** la proporción $1-\alpha$ de la distribución está incluida entre los extremos de la muestra.

La tabla A.19 proporciona los tamaños de la muestra requeridos para los valores seleccionados de γ y $1-\alpha$. Por ejemplo, cuando $\gamma = 0.01$ y $1-\alpha = 0.95$, debemos seleccionar una muestra aleatoria de tamaño $n = 130$ para tener 99% de confianza en que al menos 95% de la distribución de mediciones está incluido entre los extremos de la muestra.

En vez de determinar un tamaño muestral n tal que una proporción específica de mediciones esté contenida entre los extremos de la muestra, en muchos procesos industriales es deseable determinar un tamaño de la muestra tal que una proporción fija de la población caiga por debajo de la observación más grande (o por arriba de la más pequeña) de la muestra. Tales límites se denominan límites de tolerancia unilaterales.

Límites de tolerancia unilaterales

Para cualquier distribución de mediciones un límite de tolerancia unilateral se determina mediante la observación más pequeña (o más grande) en una muestra de tamaño n, donde n se determina de manera que se pueda asegurar con $100(1-\gamma)\%$ de confianza que **al menos** la proporción $1-\alpha$ de la distribución excederá a la observación más pequeña (menor que la mayor) de la muestra.

La tabla A.20 muestra los tamaños de la muestra requeridos, correspondientes a valores seleccionados de γ y $1-\alpha$. De aquí, cuando $\gamma = 0.05$ y $1-\alpha = 0.70$, debemos elegir una muestra de tamaño $n = 9$ para tener 95% de confianza en que 70% de nuestra distribución de mediciones excederá la observación más pequeña de la muestra.

16.8. Coeficiente de correlación de rango

En el capítulo 11 utilizamos el coeficiente de correlación muestral r para medir el coeficiente de correlación poblacional ρ, la relación lineal entre dos variables continuas X y Y. Si los rangos 1, 2,..., n se asignan a las observaciones x en orden de magnitud y de manera similar a las observaciones y, y si estos rangos se sustituyen después con los valores numéricos reales en la fórmula para el coeficiente de correlación del capítulo 11, obtenemos el equivalente no paramétrico del coeficiente de correlación convencional. Un coeficiente de correlación calculado de esta forma se conoce como **coeficiente de correlación de rangos de Spearman** y se denota con r_s. Cuando no hay empates entre ambos conjuntos de mediciones la fórmula para r_s se reduce a una expresión mucho más simple que incluye las diferencias d_i entre los rangos asignados a los n pares de x y y que establecemos ahora.

Coeficiente de correlación de rangos

Una medida no paramétrica de la asociación entre dos variables X y Y es dada por el **coeficiente de correlación de rango**

$$r_s = 1 - \frac{6}{n(n^2 - 1)} \sum_{i=1}^{n} d_i^2,$$

donde d_i es la diferencia entre los rangos asignados a x_i y y_i, y n es el número de pares de datos.

En la práctica, la fórmula anterior también se usa cuando hay empates entre las observaciones x o y. Los rangos para observaciones empatadas se asignan de la misma manera que en la prueba de rango con signo al promediar los rangos que se habrían asignado si las observaciones fueran distinguibles.

El valor de r_s por lo general se acercará al valor que se obtiene al calcular r con base en mediciones numéricas y se interpreta de forma muy similar. Como antes, el valor de r_s irá de -1 a $+1$. Un valor de $+1$ o -1 indica una asociación perfecta entre X y Y; el signo más ocurre para rangos idénticos y el signo menos para rangos inversos. Cuando r_s se acerca a cero, se concluye que las variables no están correlacionadas.

17. CONTROL ESTADÍSTICO DE CALIDAD

17.1.Introducción

La idea de usar técnicas de muestreo y análisis estadístico en un entorno de producción tuvo sus comienzos en la década de 1920. El objetivo de este concepto tan exitoso es reducir de manera sistemática la variabilidad y el aislamiento asociados con las fuentes de dificultades *durante la producción*. En 1924 Walter A. Shewhart, de la empresa Bell Telephone Laboratories, desarrolló el concepto de gráfica de control. Sin embargo, fue hasta la Segunda Guerra Mundial cuando se generalizó el uso de este tipo de gráficas debido a la importancia que durante ese periodo tuvo el mantenimiento de la calidad en los procesos de producción. En las décadas de 1950 y 1960 el desarrollo del control de calidad y el área general de seguridad de la calidad crecieron con rapidez, en particular con el surgimiento del programa espacial en Estados Unidos. En Japón hubo un amplio y exitoso uso del control de calidad gracias a los esfuerzos de W. Edwards Deming, quien trabajó como consultor en Japón después de la Segunda Guerra Mundial. El control de calidad ha sido, y es, un elemento importante en el desarrollo de la industria y la economía de Japón.

El control de calidad está recibiendo cada vez más atención como una herramienta de administración en la cual se observan y evalúan las características importantes de un producto en comparación con algún tipo de estándar. Los diversos procedimientos en el control de calidad implican un uso considerable de los procedimientos de muestreo y los principios estadísticos expuestos en capítulos anteriores. Los principales usuarios del control de calidad son, por supuesto, las corporaciones industriales. Es evidente que un programa eficaz de control de calidad mejora la calidad del artículo que se produce y aumenta las utilidades. Esto es particularmente cierto en la actualidad, pues los productos se fabrican en volúmenes altos. Antes de que surgiera el movimiento hacia los métodos de control de calidad, a menudo ésta se veía afectada debido a la falta de eficiencia, lo cual, por supuesto, incrementaba los costos.

17.2.Naturaleza de los límites de control

Las ideas fundamentales en las que se basan las gráficas de control son similares en estructura a la prueba de hipótesis. Los límites de control se establecen para controlar la probabilidad de cometer el error de concluir que el proceso está fuera de control, cuando de hecho no lo está. Esto corresponde a la probabilidad de cometer un error tipo I si probáramos la hipótesis nula de que el proceso está bajo control. Por otro lado, debemos estar atentos al error del segundo tipo, es decir, el de no encontrar el proceso fuera de control cuando de hecho sí lo está (error tipo II). De esta manera, la elección de los límites de control es similar a la elección de una región crítica.

Como en el caso de la prueba de hipótesis, el tamaño de la muestra en cada punto es importante. La elección del tamaño de la muestra depende en gran medida de la sensibilidad o potencia de detección del estado fuera de control. En esta aplicación, el concepto de *potencia* es muy similar al de la situación de la prueba de hipótesis. Queda claro que cuanto más grande sea la muestra en cada periodo, más rápida será la detección de un proceso fuera de control. En cierto sentido los límites de control en realidad definen lo que el usuario considera como estar *bajo control.* En otras palabras, la amplitud dada por los límites de control debe depender en cierto sentido de la variabilidad del proceso. Como resultado, el cálculo de los límites de control dependerá de manera natural de los datos que se tomen de los resultados del proceso. De esta forma, cualquier aplicación del control de calidad debe comenzar con el cálculo de una muestra o conjunto de muestras preliminar, que establecerá tanto la línea central como los límites del control de calidad.

17.3.Propósitos de la grafica de control

Un propósito evidente de la gráfica de control es la vigilancia del proceso, o sea determinar si es o no necesario realizar cambios. Además, la constante y sistemática obtención de datos a menudo permite a la administración evaluar la capacidad del proceso. Es evidente que, si una sola característica de desempeño es importante, el muestreo y la estimación continuos de la media y la desviación estándar de esa característica de desempeño ofrecen la actualización de lo que el proceso puede hacer en términos de desempeño promedio y variación aleatoria. Esto es valioso incluso cuando el proceso permanece bajo control durante periodos largos. La estructura sistemática y formal de la gráfica de control a menudo puede prevenir una reacción desmesurada ante cambios que representen sólo fluctuaciones aleatorias. Obviamente, en muchas situaciones los cambios realizados por una reacción desmesurada pueden crear graves problemas que son difíciles de resolver.

Las características de calidad de las gráficas de control por lo general caen en *dos* categorías: **variables** y **atributos**. Como resultado, los tipos de gráficas de control con frecuencia tienen las mismas clasificaciones. En el caso de la gráfica de los tipos de variables, la característica suele ser una medida sobre un continuo, como el diámetro o el peso. En el caso de la gráfica de atributos, lo que refleja la característica es si el producto individual se *ajusta a las especificaciones* (si está o no defectuoso). Las aplicaciones para estas dos situaciones distintas son evidentes.

En el caso de la gráfica de variables se debe ejercer control sobre la tendencia central y la variabilidad. Lo que a un analista de control de calidad le debe preocupar es si existe o no, *en promedio,* un cambio en los valores de la característica de desempeño. Además, siempre habrá interés por saber si algún cambio en las condiciones del proceso provoca que disminuya la precisión, es decir, que aumente la variabilidad. Para manejar estos dos conceptos es esencial utilizar gráficas de control separadas. La tendencia central es controlada por la *gráfica* $\bar{X}$, donde las medias de muestras relativamente pequeñas se dibujan en la gráfica de control. La variabilidad alrededor de la media se controla mediante el *rango* en la muestra, o la *desviación estándar de la* muestra. En el caso de muestreo de atributos a menudo la cantidad que se grafica es la *proporción de artículos defectuosos* de una muestra. En la siguiente sección analizamos el desarrollo de gráficas de control para los tipos de variables de las características del desempeño.

17.4.Graficas de control para las variables

Un ejemplo es una forma relativamente sencilla de explicar los rudimentos de la gráfica $\bar{X}$ para variables. Suponga que en un proceso de fabricación de cierta parte de un motor se deben utilizar las gráficas de control de calidad. Suponga también que la media del proceso es $\mu = 50$ mm y que la desviación estándar es $\sigma = 0.01$ mm. Imagine que se toman muestras en grupos de 5 cada hora y que los valores de la *media muestral* $\bar{X}$ se registran y grafican como en la figura 17.2. Los límites para las gráficas $\bar{X}$ se basan en la desviación estándar de la variable aleatoria $\bar{X}$. Sabemos, a partir de lo expuesto en el capítulo 8, que para el promedio de observaciones independientes en una muestra de tamaño n,

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}},$$

donde σ es la desviación estándar de una observación individual. Los límites de control están diseñados para dar como resultado una pequeña probabilidad de que un valor dado de $\bar{X}$ esté fuera de los límites dado que, en realidad, el proceso está bajo control, es decir, $\mu = 50$. Si recurrimos al teorema del límite central, tendremos que, en las condiciones en las que el proceso está controlado,

$$\bar{X} \sim N\left(50, \frac{0.01}{\sqrt{5}}\right).$$

Como resultado, $100(1-\alpha)\%$ de los valores $\bar{X}$ cae dentro de los límites cuando el proceso está bajo control si utilizamos los límites

$$\text{LCI} = \mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 50 - z_{\alpha/2}(0.0045), \qquad \text{LCS} = \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 50 + z_{\alpha/2}(0.0045).$$

Aquí LCI y LCS representan el límite de control inferior y el límite de control superior, respectivamente. Con frecuencia las gráficas $\bar{X}$ se basan en límites denominados "tres-sigma", refiriéndonos, por supuesto, a $z_{\alpha/2} = 3$ y a límites que se convierten en

$$\mu \pm 3\frac{\sigma}{\sqrt{n}}.$$

En nuestro ejemplo, los límites superior e inferior son

$$\text{LCI} = 50 - 3(0.0045) = 49.9865, \qquad \text{LCS} = 50 + 3(0.0045) = 50.0135.$$

Por consiguiente, si vemos la estructura de los límites 3σ desde el punto de vista de la prueba de hipótesis para un punto muestral dado, encontraremos que hay una probabilidad

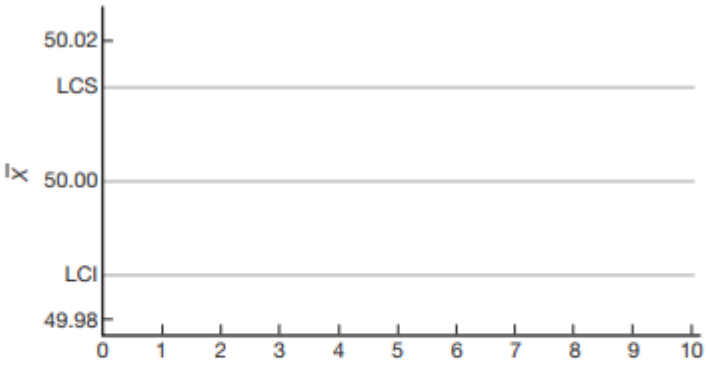


Figura 17.2: Los límites de control 3σ para el ejemplo de la parte del motor.

de 0.0026 de que el valor $\bar{X}$ caiga fuera de los límites de control, dado que el proceso está bajo control. Ésta es la probabilidad de que el analista determine *de manera errónea* que el proceso está fuera de control (véase la tabla A.3).

El ejemplo anterior no sólo ilustra la gráfica $\bar{X}$ para las variables, también proporciona al lector una idea general de la naturaleza de las gráficas de control. La línea central por lo general refleja el valor ideal de un parámetro importante. Los límites de control se establecen a partir del conocimiento de las propiedades de muestreo del estadístico que estima el parámetro en cuestión. Con mucha frecuencia implican un múltiplo de la desviación estándar del estadístico. Se ha generalizado el uso de límites 3σ. En el caso de la gráfica $\bar{X}$ que se presenta aquí, el teorema del límite central brinda al usuario una buena aproximación de la probabilidad de determinar de forma errónea que el proceso está fuera de control. En general, sin embargo, es probable que el usuario no confíe en la normalidad del estadístico sobre la línea central. Lo anterior podría dar como resultado que no se conozca la probabilidad exacta de cometer un "error tipo I". A pesar de esto se ha vuelto muy común utilizar los límites $k\sigma$. Aunque los límites 3σ se utilizan ampliamente, en ocasiones el usuario utilizará otro método. Cuando es importante detectar de forma rápida una situación fuera de control podría ser apropiado utilizar un múltiplo menor de σ. Si se toman en cuenta los costos de producción, cabe señalar que permitir que un proceso continúe funcionando fuera de control, incluso por periodos cortos, puede resultar más costoso que invertir en la investigación y corrección de las causas de la pérdida del control en el proceso. En este caso es evidente que los límites apropiados son los límites de control que son más estrictos que los límites 3σ.

Subgrupos racionales

Los valores de la muestra que se utilizan para el control de calidad se dividen en subgrupos, en los que una *muestra* representa un subgrupo. Como antes indicamos, el orden en el tiempo de producción es en realidad una base natural para la selección de los subgrupos. Podríamos considerar el esfuerzo de control de calidad de manera muy simple como 1) muestreo, 2) detección de un estado fuera de control y 3) búsqueda de las causas atribuibles que puedan ocurrir con el tiempo. Tal vez parezca que la selección de la base para estos grupos muestrales es muy sencilla, pero la elección de estos subgrupos de información muestral podría tener un efecto importante en el éxito del programa de control de calidad. Estos subgrupos con frecuencia se denominan **subgrupos racionales**. En general, si el analista está interesado en detectar un *cambio de ubicación,* se considera que los subgrupos se deben elegir de manera que la variabilidad dentro del subgrupo sea pequeña, y de manera que haya mayores posibilidades de detectar las causas atribuibles, si se presentaran. Así, deseamos elegir los subgrupos de forma que se maximice la variabilidad entre subgrupos. Por ejemplo, un método razonable es elegir unidades en un subgrupo que se producen de forma cercana en el tiempo. Por otro lado, las gráficas de control a menudo se utilizan para controlar la variabilidad, en cuyo caso el estadístico de desempeño es la *variabilidad dentro de la muestra.* Por consiguiente, es más importante elegir los subgrupos racionales para maximizar la variabilidad dentro de la muestra. En este caso las observaciones en los subgrupos se deberían comportar más como una muestra aleatoria y la variabilidad dentro de las muestras necesita ser una descripción de la variabilidad del proceso.

Es importante señalar que las gráficas de control sobre la variabilidad se deben establecer antes de construir gráficas sobre el centro de ubicación (digamos, gráficas $\bar{X}$). Cualquier gráfica de control sobre el centro de ubicación en realidad dependerá de la variabilidad. Por ejemplo, vimos un ejemplo de la gráfica de tendencia central y ésta depende de σ. En las secciones que siguen se analizará un estimado de σ a partir de los datos.

17.5.Graficas de control para artículos defectuosos (uso del modelo de Poisson)

En el procedimiento anterior supusimos que el artículo bajo consideración es uno que está defectuoso, es decir, que no funciona, o uno que no tiene defecto, en cuyo caso el artículo funciona y, por lo tanto, es aceptable para el consumidor. En muchas situaciones este método del artículo "defectuoso o no" es demasiado simplista. Las unidades pueden contener defectos o no cumplir con las especificaciones, y aun así funcionar bastante bien para el consumidor. En realidad, en este caso sería importante ejercer control sobre el *número de defectos* o *número de artículos que no cumplen las especificaciones*. Este tipo de control de calidad tiene aplicación cuando las unidades no son simplistas ni grandes. Por ejemplo, el número de defectos puede ser muy útil como objeto de control cuando el artículo o unidad es, digamos, una computadora personal. Otro ejemplo es una unidad definida por 50 pies de tubería fabricada, donde el número de soldaduras defectuosas es el objeto del control de calidad; el número de defectos en 50 pies de alfombra fabricada o el número de "burbujas" en una hoja grande de vidrio fabricado.

A partir de lo aquí descrito queda claro que en este caso no es apropiada la distribución binomial. El número total de artículos que no cumplen las especificaciones en una unidad o el número promedio por unidad se podría usar como la medida para la gráfica de control. A menudo se supone que el número de artículos que no cumplen las especificaciones en una muestra tiene una distribución de Poisson. A este tipo de gráfica con frecuencia se le llama **gráfica *C***.

Suponga que el número de defectos X en una unidad de producto tiene una distribución de Poisson con parámetro λ. (Aquí $t = 1$ para el modelo de Poisson). Recuerde que para la distribución de Poisson,

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad x = 0, 1, 2, \ldots.$$

Aquí, la variable aleatoria X es el número de artículos que no cumplen las especificaciones. En el capítulo 5 vimos que tanto la media como la varianza de la variable aleatoria de Poisson son λ. Por consiguiente, si la gráfica de control de calidad se estructurara de acuerdo con los límites 3σ acostumbrados, si conociéramos λ tendríamos,

Ejemplo 17.5: La tabla 17.4 representa el número de defectos en 20 muestras sucesivas de rollos de hoja metálica, cada uno con 100 pies de largo. Para controlar el número de defectos en tales muestras se debe desarrollar una gráfica de control a partir de estos datos preliminares. El estimado del parámetro de Poisson λ es dado por $\hat{\lambda} = 5.95$. Como resultado, los límites de control sugeridos por estos datos preliminares son

$$\text{LCS} = \hat{\lambda} + 3\sqrt{\hat{\lambda}} = 13.2678 \quad \text{y} \quad \text{LCI} = \hat{\lambda} - 3\sqrt{\hat{\lambda}} = -1.3678,$$

donde LCI se iguala a cero.

Tabla 17.4: Datos para el ejemplo 17.5; el control implica el número de defectos en rollos de hojas metálicas

Número de muestra	Número de defectos	Número de muestra	Número de defectos
1	8	11	3
2	7	12	7
3	5	13	5
4	4	14	9
5	4	15	7
6	7	16	7
7	6	17	8
8	4	18	6
9	5	19	7
10	6	20	4
			Prom. 5.95

La figura 17.9 presenta una gráfica de los datos preliminares con los límites de control.

La tabla 17.5 incluye datos adicionales tomados del proceso de producción. Para cada muestra se inspeccionó la unidad en la que se basó la gráfica, a saber, 100 pies del metal. Se incluye la información de 20 muestras. La figura 17.10 muestra una gráfica de los datos adicionales de producción. Es evidente que el proceso está bajo control, o al menos lo estaba en el periodo en el que se tomaron los datos. ∎

Tabla 17.5: Datos adicionales del proceso de producción del ejemplo 17.5

Número de muestra	Número de defectos	Número de muestra	Número de defectos
1	3	11	7
2	5	12	5
3	8	13	9
4	5	14	4
5	8	15	6
6	4	16	5
7	3	17	3
8	6	18	2
9	5	19	1
10	2	20	6

En el ejemplo 17.5 dejamos muy claro que la unidad de muestreo o de inspección son 100 pies de metal. En muchos casos en los que el artículo es específico, como en el caso de una computadora personal o el de un tipo específico de dispositivo electrónico, la unidad de inspección podría ser un *conjunto de artículos*. Por ejemplo, el analista decide utilizar 10 computadoras en cada subgrupo y de esta forma observar un conteo del número total de defectos encontrados. Por consiguiente, la muestra preliminar para construir la gráfica de control implica utilizar varias muestras, cada una de 10 computadoras. La elección del tamaño de la muestra puede depender de muchos factores. A menudo deseamos un tamaño de la muestra que asegure un LCI positivo.

El analista podría utilizar el número promedio de defectos por unidad de muestreo como la medida básica de la gráfica de control. Por ejemplo, para el caso de la compu-

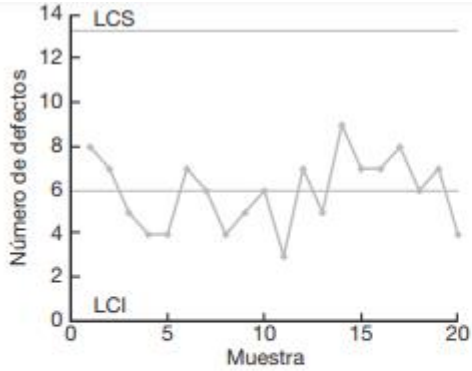


Figura 17.9: Datos preliminares representados en la gráfica de control para el ejemplo 17.5.

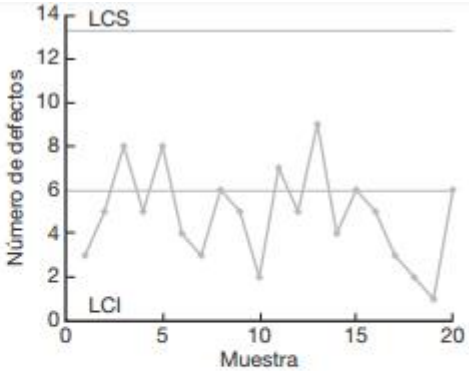


Figura 17.10: Datos adicionales de producción para el ejemplo 17.5.

tadora personal, sea la variable aleatoria el número total de defectos

$$U = \frac{\text{número total de defectos}}{n}$$

que se mide para cada muestra de, digamos, $n = 10$. Si suponemos que el número de defectos por unidad de muestreo es de Poisson con parámetro λ, podemos utilizar el método de las funciones generadoras de momento para demostrar que U es una variable aleatoria de Poisson (véase el ejercicio de repaso 17.1). De esta manera, la gráfica de control para esta situación se caracteriza por lo siguiente:

$$\text{LCS} = \bar{U} + 3\sqrt{\frac{\bar{U}}{n}}, \qquad \text{línea central} = \bar{U}, \qquad \text{LCI} = \bar{U} - 3\sqrt{\frac{\bar{U}}{n}}.$$

Aquí, desde luego, $\bar{U}$ es el promedio de los valores U en el conjunto de datos preliminares o base. El término $\bar{U}/n$ se deriva del resultado que

$$E(U) = \lambda, \qquad \text{Var}(U) = \frac{\lambda}{n},$$

y por ello U es un estimado no sesgado de $E(U) = \lambda$ y U/n es un estimado no sesgado de $\text{Var}(U) = \lambda/n$. Este tipo de gráfica de control a menudo se denomina **gráfica *U***.

En esta sección basamos toda la explicación de las gráficas de control en el modelo de probabilidad de Poisson. Este modelo se ha utilizado en combinación con el concepto 3σ. Como explicamos antes en este capítulo, el concepto de límites 3σ tiene sus raíces en la aproximación normal, aunque muchos usuarios consideran que el concepto funciona bien como herramienta pragmática incluso si la normalidad no es siquiera aproximadamente correcta. La dificultad, desde luego, radica en el hecho de que, en ausencia de normalidad, no es posible controlar la probabilidad de una especificación incorrecta de un estado fuera de control. En el caso del modelo de Poisson, cuando λ es pequeña la distribución es bastante asimétrica, una condición que puede producir resultados indeseables si se utiliza el método 3σ.

17.6.Graficas de control de cusum

La desventaja de las gráficas de control similares a las de Shewhart, que se explicaron y ejemplificaron en las secciones anteriores, radica en su incapacidad para detectar pequeños cambios en la media. Un mecanismo de control de calidad que ha recibido mucha atención en la literatura estadística y que se ha utilizado extensamente en la industria es la **gráfica de suma acumulada (cusum)**. El método de la gráfica de suma acumulada es sencillo y, por lo tanto, atractivo. Para el lector debe ser evidente por qué es más sensible a pequeños cambios en la media. Considere una gráfica de control para la media con un nivel de referencia establecido en el valor W. Considere las observaciones particulares $X_1, X_2, \ldots, X_r$. Las primeras cusum r son

$$\begin{aligned} S_1 &= X_1 - W \\ S_2 &= S_1 + (X_2 - W) \\ S_3 &= S_2 + (X_3 - W) \\ &\vdots \\ S_r &= S_{r-1} + (X_r - W). \end{aligned}$$

Es evidente que la cusum es simplemente la acumulación de las diferencias del nivel de referencia. Es decir,

$$S_k = \sum_{i=1}^{k} (X_i - W), \qquad k = 1, 2, \ldots.$$

La gráfica cusum es, entonces, una gráfica de S_k contra el tiempo.

Suponga que consideramos que el nivel de referencia W es un valor aceptable de la media μ. Salta a la vista que, si no hay cambio en μ, la gráfica cusum debería ser aproximadamente horizontal, con algunas fluctuaciones menores balanceadas alrededor de cero. Ahora, si sólo hay un cambio moderado en la media, debe resultar un cambio más o menos grande en la *pendiente* de la gráfica cusum, dado que cada nueva observación tiene la probabilidad de contribuir a un cambio y la medida que se grafica acumula esos cambios. Desde luego, la señal de que la media ha cambiado reside en la naturaleza de la pendiente de la gráfica cusum. El objetivo de la gráfica es detectar cambios que se alejan del nivel de referencia. Una pendiente diferente de cero (en cualquier dirección) representa un cambio a partir del nivel de referencia. Una pendiente positiva indica un aumento en la media por arriba del nivel de referencia, en tanto que una pendiente negativa señala una disminución.

Las gráficas cusum a menudo se diseñan con un *nivel de calidad aceptable* definido (NCA) y un *nivel de calidad rechazable* (NCR) preestablecido por el usuario. Ambos representan valores de la media. Se podría considerar que éstos desempeñan papeles similares a los de las medias nula y alternativa en la prueba de hipótesis. Considere una situación en la que el analista desea detectar un aumento en el valor de la media del proceso. Usaremos la notación μ_0 para NCA y μ_1 para NCR, y $\mu_1 > \mu_0$. El nivel de referencia se fija ahora en

$$W = \frac{\mu_0 + \mu_1}{2}.$$

Los valores de S_r $(r = 1, 2,...)$ tendrán una pendiente negativa si la media del proceso está en μ_0 y una pendiente positiva si la media del proceso está en μ_1.

Regla de decisión para las gráficas cusum

Como antes se expuso, la pendiente de la gráfica cusum proporciona la señal de acción para el analista de control de calidad. La regla de decisión exige tomar medidas si, en el r-ésimo periodo de muestreo,

$$d_r > h,$$

donde h es un valor preestablecido que se denomina **longitud del intervalo de decisión** y

$$d_r = S_r - \min_{1 \le i \le r-1} S_i.$$

En otras palabras, se toman medidas si los datos revelan que el valor de la cusum real excede en una cantidad específica al valor previo de la cusum más pequeño.

Una modificación en la mecánica que se describió antes facilita el uso del método. Describimos un procedimiento que grafica las cusum y calcula las diferencias. Una modificación simple implica graficar las diferencias de manera directa y permitir la verificación contra el intervalo de decisión. La expresión general para d_r es muy sencilla. Para el procedimiento de cusum, con el que se detectan aumentos en la media,

$$d_r = \text{máx}[0, d_{r-1} + (X_r - W)].$$

La elección del valor de h es, por supuesto, muy importante. En este libro no se proporcionan los detalles que aparecen en la literatura que trata de esta elección. Para una exposición más completa se remite al lector a Ewan y Kemp, 1960, y a Montgomery, 2000b (véase la bibliografía). Una consideración importante es la **longitud esperada de la corrida**. De manera ideal, la longitud esperada de la corrida es bastante grande bajo $\mu = \mu_0$ y muy pequeña cuando $\mu = \mu_1$.

BIBLIOGRAFÍA

JEFFREY, R.C. (1992)., Probability and the Art of Judgment, Cambridge University Press. pp. 54-55. ISBN 0-521-39459-7

WOLFOWITZ, J. (1969): 'Reflections on the future of mathematical statistics'. en R. c. Bose et al. (eds.) "Essays in Probability and Sraristics". University of North Carolina Press. Chapel Hill.

TUKEY, J.W. (1962): 'The future of Data Analysis'. Annals of Mathematical Statistics, 33, 1-67.

EDWIN THOMPSON JAYNES (1996).. Probability Theory: The Logic of Science. Preprint: Washington University, (PDF) (en inglés)

RAFAEL DÍAZ (2000). Introducción a la Probabilidad, los Procesos Estocásticos y la Estadística en Ingenie ría. Escuela deIngeniería Eléctrica. Universidad Central de Venezuela.

NALCÁZAR, Marco, (2002), Unidades para Producir Medios Instruccionales en Educación, SUÁREZ, Mario Ed. Graficolor, Ibarra, Ecuador.

DAZA, Jorge, (2006), Estadística Aplicada con Microsoft Excel, Grupo Editorial Megabyte, Lima, Perú.

GOVINDEN, Lincoyán, (1985), Introducción a la Estadística, Ed. McGraw Hill. Interamericana Editores. S.A., Bogotá, Colombia.

JOHNSON, Robert, (2003), Estadística Elemental, Ed. Math Learning, Ed. Tercera, México DF. KUBY, Patricia.

KAZMIER, J. Leonard, (1990). Estadística Aplicada a la Administración y la Economía, Ed. McGrawHill, Ed. Segunda, Bogotá, Colombia.

LIND, Marchal, (2005), Estadística Aplicada a los Negocios y a la Economía, Ed. McGraw- Hill, MASON Ed. Décima., México DF.

MARTINEZ, Bencardino, (1981), Estadística Comercial, Ed. Norma, Bogotá, Colombia.

MORENO, Francis, (1993), Estadística Inferencial, Universidad Particular de Loja, Loja, Ecuador.

SÁNCHEZ, Jesús, (2007), Introducción a la Estadística Empresarial, Madrid, España.

SALTOS, Héctor, (1986), Estadística de Inferencia, Ed. Pío XII, Ambato, Ecuador.

SHAO, Stephen, (1980), Estadística para Economistas y Administradores de Empresas, Ed. Herrero Hnos, México DF.

SPIEGEL, Murray, (2000), Estadística, Serie de Compendios Schaum, Ed. McGraw-Hill, México.

SUÁREZ, Mario, (2004), Interaprendizaje Holístico de Matemática, Ed. Gráficas Planeta, Ibarra,

Ecuador.

STEVENSON, William, (1981), Estadística para Administración y Economía, Ed. Harla S.A de C.V.

México D.F.

WEBSTER, Allen, (2000), Estadística Aplicada a los Negocios y a la Economía, Ed. McGraw Hill.

Kolmogorov AN. 1933 Berlin: Springer-Verlag;. (Traducido al inglés: Morrison N. Foundations of the Theory of Probability. New York: Chelsea; 1956).

Martín Pliego FJ, Ruiz-Maya L 1997. Estadística I: Probabilidad. Madrid: Editorial AC;

Meyer PL (1986).. Probabilidad y Aplicaciones Estadísticas. México: Addison-Wesley Iberoamericana.

Peña D. Modelos y métodos (1993). 1. Fundamentos. Madrid: Alianza Universidad Textos;

Katz DL (1997). Epidemiology, Biostatistics and Preventive Medicine Review. USA: W.B. Saunders Company.

Doménech JM. (1997) Métodos Estadísticos en Ciencias de la Salud. Barcelona: Signo; Hospital Ramón y Cajal. Material docente de la unidad de bioestadística clínica. Disponible en: http://www.hrc.es/bioest/M_docente.html

Bartlett, M. S. y Kendall, D. G. (1946). "The Statistical Analysis of Variance Heterogeneity and Logarithmic Transformation", Journal of the Royal Statistical Society, Ser. B. 8, 128-138.

Bowker, A. H. y Lieberman, G. J. (1972). Engineering Statistics, 2.a, ed. Upper Saddle River, N.J.: Prentice Hall.

Box, G. E. P. (1988). "Signal to Noise Ratios, Performance Criteria and Transformations (with discussion)", Technometrics, 30, 1-17.

Box, G. E. P. y Fung, C. A. (1986). "Studies in Quality Improvement: Minimizing Transmitted Variation by Parameter Design", Informe 8. University of Wisconsin-Madison, Center for Quality and Productivity Improvement.

Box, G. E. P., Hunter, W. G. y Hunter, J. S. (1978). Statistics for Experimenters. Nueva York: John Wiley & Sons.

Brownlee, K. A. (1984). Statistical Theory and Methodology: In Science and Engineering, 2.a, ed., Nueva York: John Wiley & Sons.

Carroll, R. J. y Ruppert, D. (1988). Transformation and Weighting in Regression. Nueva York: Chapman y Hall.

Chatterjee, S., Hadi, A. S. y Price, B. (1999). Regression Analysis by Example, 3.a, ed., Nueva York: John Wiley & Sons.

Cook, R. D. y Weisberg, S. (1982). Residuals and Influence in Regression. Nueva York: Chapman y Hall.

Daniel, C. y Wood, F. S. (1999). Fitting Equations to Data: Computer Analysis of Multifactor Data, 2.a, ed., Nueva York: John Wiley & Sons.

Daniel, W. W. (1989). Applied Nonparametric Statistics, 2.a, ed. Belmont, Calif.: Wadsworth Publishing Company.

Devore, J. L. (2003). Probability and Statistics for Engineering and the Sciences, 6.a, ed., Belmont, Calif: Duxbury Press.

Dixon, W. J. (1983). Introduction to Statistical Analysis, 4.a, ed., Nueva York: McGraw-Hill.

Draper, N. R. y Smith, H. (1998). Applied Regression Analysis, 3.a, ed., Nueva York: John Wiley & Sons.

Duncan, A. (1986). Quality Control and Industrial Statistics, 5.a, ed., Homewood, Ill.: Irwin.

Dyer, D. D., y Keating, J. P. (1980). "On the Determination of Critical Values for Bartlett's Test", Journal of the American Statistical Association, 75, 313-319.

Ewan, W. D. y Kemp, K. W. (1960). "Sampling Inspection of Continuous Processes with No Autocorrelation between Successive Results", Biometrika, 47, 363-380.

Geary, R. C. (1947). "Testing for Normality", Biometrika, 34, 209-242.

Gunst, R. F. y Mason, R. L. (1980). Regression Analysis and Its Application: A Data-Oriented Approach. Nueva York: Marcel Dekker.

Guttman, I., Wilks, S. S. y Hunter, J. S. (1971). Introductory Engineering Statistics. Nueva York: John Wiley & Sons.

Harville, D. A. (1 977). "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems", Journal of the American Statistical Association, 72, 320-338.

Hicks, C. R. y Turner, K. V. (1999). Fundamental Concepts in the Design of Experiments, 5.a, ed., Oxford: Oxford University Press.

Hoaglin, D. C., Mosteller, F. y Tukey, J. W. (1991). Fundamentals of Exploratory Analysis of Variance. Nueva York: John Wiley & Sons.

Hocking, R. R. (1976). "The Analysis and Selection of Variables in Linear Regression", Biometrics, 32, 1-49.

Hodges, J. L. y Lehmann, E. L. (2005). Basic Concepts of Probability and Statistics, 2.a, ed. Filadelfi a: Society for Industrial and Applied Mathematics.

26 Hoerl, A. E. y Wennard, R. W. (1970). "Ridge Regression: Applications to Nonorthogonal Problems",

Technometrics, 12, 55-67.

27 Hogg, R. V. y Ledolter, J. (1992). Applied Statistics for Engineers and Physical Scientists, 2.a, ed., Upper Saddle River, N.J.: Prentice Hall.

Hogg, R. V., McKean, J. W. y Craig, A. (2005). Introduction to Mathematical Statistics, 6.a, ed., Upper Saddle River, N.J.: Prentice Hall.

Hollander, M. y Wolfe, D. (1999). Nonparametric Statistical Methods. Nueva York: John Wiley & Sons.

Johnson, N. L. y Leone, F. C. (1977). Statistics and Experimental Design: In Engineering and the Physical Sciences, 2.a, ed., Vols. I y II, Nueva York: John Wiley & Sons.

Kackar, R. (1985). "Off-Line Quality Control, Parameter Design, and the Taguchi Methods", Journal of Quality Technology, 17, 176-188.

Koopmans, L. H. (1987). An Introduction to Contemporary Statistics, 2.a, ed., Boston: Duxbury Press.

Kutner, M. H., Nachtsheim, C. J., Neter, J. y Li, W. (2004). Applied Linear Regression Models, 5.a, ed., Nueva York: McGraw-Hill/Irwin

APÉNDICE A TABLAS Y DEMOSTRACIONES ESTADÍSTICAS

Copy protected with Online-PDF-No-Copy.com